Proceedings

THIRTY-EIGHTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL AND COMPUTING

October 4 - 6, 2000

Allerton House, Monticello, Illinois
Sponsored by the
Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering of the
University of Illinois at Urbana-Champaign

PROCEEDINGS

THIRTY-EIGHTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL, AND COMPUTING

VOLUME I

R.S. Sreenivas
Douglas L. Jones
Conference Co-Chairs

Conference held
October 4 - 6, 2000
Allerton House
Monticello, Illinois

Sponsored by
The Coordinated Science Laboratory
and
The Department of Electrical and Computer Engineering
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

621.3811
AL5
2000:1

FOREWORD

These Proceedings contain papers that were presented at the Thirty-eighth Annual Allerton Conference on Communication, Control, and Computing which was held at Allerton House, Monticello, Illinois, on October 4 through October 6, 2000. As in past years, the Conference was sponsored by the Coordinated Science Laboratory and the Department of Electrical and Computer Engineering of the University of Illinois at Urbana-Champaign.

Ninety-five invited papers and sixty-seven contributed papers were presented at the Conference in twenty-nine sessions. We thank all the authors and particularly the speakers for contributing to the success of the Conference. We are also grateful to our colleagues at the University of Illinois at Urbana-Champaign: Professors D. Baron, C. Beck, R.E. Blahut, N. Boston, F. Bullo, G. Dullerud, C. Hadjicostis, B. Hajek, T.S. Huang, R. Koetter, P.R. Kumar, S.P. Meyn, P. Moulin, D. Sarwate, A. Singer, R. Srikant, P.G. Voulgaris, and V.V. Veeravalli for organizing many of the invited sessions, reviewing the contributed papers, chairing sessions, and helping in many other ways in the smooth running of the Conference. Professors R.J. Barton of the University of Houston, G. Caire of EURECOM, E.K.P. Chong of Purdue University, D. Forney of the Massachusetts Institute of Technology, D.L. Goeckel of the University of Massachusetts, Amherst, A. Goldsmith of Stanford University, S.G. Henderson of the University of Michigan, M. Médard of the Massachusetts Institute of Technology, A.M. Sayeed of the University of Wisconsin, Madison, and X. Wang of Texas A&M University chaired contributed and invited sessions at the Conference, and for this we thank them.

The plenary speaker at the Conference was Robert J. McEliece of California Institute of Technology, who gave an illuminating lecture titled "Achieving the Shannon Limit: a Progress Report." We are grateful to him for his excellent presentation. We are also indebted to the entertainment from "No Strings Attached," for giving us a wonderful evening of a capella music.

We thank Mr. Jim Gortner and his staff at Allerton House for cheerfully accommodating our Conference and for their constant attention to our needs. It is also a great pleasure to thank all the staff members of the Coordinated Science Laboratory who so ably handled all the practical aspects of organizing the Conference. In particular, we wish to mention Mr. Steven Williams and Ms. Susan Baxley who handled all the business affairs, and Mr. Dan Jordan who handled the sales and shipping of these Proceedings. Most of all, however, we would like to thank our Conference Secretary, Mrs. Becky Lonberger, who with the able assistance of Mrs. Francie Bridges and Mrs. Terri Hovde worked all year round in handling announcements, calls for papers, program preparation, Allerton House arrangements, registration, and the final editing of these Proceedings. The lion's share of the credit for the success of the Conference belongs to these individuals and it gives us the greatest pleasure to acknowledge this here.

R.S. Sreenivas and Douglas L. Jones

Table 1: 35 YEARS OF ALLERTON CONFERENCES

Name	Dates	Chair(s)
First Annual Allerton Conference on Circuit and System Theory	November 15-17, 1963	J. B. Cruz, Jr.
Second Annual Allerton Conference on Circuit and System Theory	September 28-30, 1964	W. R. Perkins
Third Annual Allerton Conference on Circuit and System Theory	October 20-22, 1965	M. E. VanValkenburg
Fourth Annual Allerton Conference on Circuit and System Theory	October 5-7, 1966	William R. Perkins J. B. Cruz, Jr.
Fifth Annual Allerton Conference on Circuit and System Theory	October 4-6, 1967	J. B. Cruz, Jr. T. N. Trick
Sixth Annual Allerton Conference on Circuit and System Theory	October 2-4, 1968	T. N. Trick R. T. Chien
Seventh Annual Allerton Conference on Circuit and System Theory	October 8-10, 1969	A. H. Haddad G. Metze T. N. Trick
Eighth Annual Allerton Conference on Circuit and System Theory	October 7-9, 1970	A. H. Haddad B. C. Kuo G. Metze
Ninth Annual Allerton Conference on Circuit and System Theory	October 6-8, 1971	J. B. Cruz, Jr. A. H. Haddad
Tenth Annual Allerton Conference on Circuit and System Theory	October 4-6, 1972	J. B. Cruz, Jr. C. L. Chen
Eleventh Annual Allerton Conference on Circuit and System Theory	October 3-5, 1973	W. R. Perkins P. V. Kokotovic
Twelfth Annual Allerton Conference on Circuit and System Theory	October 2-4, 1974	P. V. Kokotovic E. S. Davidson
Thirteenth Annual Allerton Conference on Circuit and System Theory	October 1-3, 1975	E. S. Davidson T. N. Trick
Fourteenth Annual Allerton Conference on Circuit and System Theory	September 29-October 1, 1976	T. N. Trick M. E. VanValkenburg
Fifteenth Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1977	F. P. Preparata M. B. Pursley
Sixteenth Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 1978	M. B. Pursley J. B. Cruz, Jr.
Seventeenth Annual Allerton Conference on Communication, Control, and Computing	October 10-12, 1979	J. B. Cruz, Jr. F. P. Preparata
Eighteenth Annual Allerton Conference on Communication, Control, and Computing	October 8-10, 1980	D. V. Sarwate W. R. Perkins

Name	Dates	Chair(s)
Nineteenth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1981	W. K. Jenkins D. V. Sarwate
Twentieth Annual Allerton Conference on Communication, Control, and Computing	October 6-8, 1982	H. V. Poor W. K. Jenkins
Twenty-First Annual Allerton Conference on Communication, Control, and Computing	October 5-7, 1983	Tamer Başar Donna J. Brown H. Vincent Poor
Twenty-Second Annual Allerton Conference on Communication, Control, and Computing	October 3-5, 1984	David C. Munson, Jr. Tamer Başar
Twenty-Third Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1985	Bruce Hajek David C. Munson, Jr.
Twenty-Fourth Annual Allerton Conference on Communication, Control, and Computing	October 1-3, 1986	Michael C. Loui Juraj V. Medanic
Twenty-Fifth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1987	P. R. Kumar Michael C. Loui
Twenty-Sixth Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1988	Mark W. Spong Bruce Hajek
Twenty-Seventh Annual Allerton Conference on Communication, Control, and Computing	September 27-29, 1989	Juraj V. Medanic P. R. Kumar
Twenty-Eighth Annual Allerton Conference on Communication, Control, and Computing	October 3-5, 1990	Donna J. Brown Kameshwar R. Poolla
Twenty-Ninth Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1991	George Cybenko C. Polychronopoulos
Thirtieth Annual Allerton Conference on Communication, Control, and Computing	September 30-October 2, 1992	Paul Van Dooren Mark W. Spong
Thirty-First Annual Allerton Conference on Communication, Control, and Computing	September 29-October 1, 1993	Dilip V. Sarwate Paul Van Dooren
Thirty-Second Annual Allerton Conference on Communication, Control, and Computing	September 28-30, 1994	W. Kenneth Jenkins Dilip V. Sarwate
Thirty-Third Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 1995	Sean Meyn Ken Zeger
Thirty-Fourth Annual Allerton Conference on Communication, Control, and Computing	October 2-4, 1996	Sean P. Meyn W. Kenneth Jenkins
Thirty-Fifth Annual Allerton Conference on Communication, Control, and Computing	September 29-October 1, 1997	Tamer Başar Alexander Vardy
Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing	September 23-25, 1998	Tamer Başar Bruce Hajek
Thirty-Seventh Annual Allerton Conference on Communication, Control, and Computing	September 22-24, 1999	Bruce Hajek R.S. Sreenivas
Thirty-Eighth Annual Allerton Conference on Communication, Control, and Computing	October 4-6, 2000	R.S. Sreenivas Douglas.L. Jones

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i

I-A: WIRELESS COMMUNICATIONS I: PHYSICAL LAYER DESIGN

Organizer: V.V. Veeravalli
(University of Illinois at Urbana-Champaign)
Chair: G. Caire
(EURECOM)

ADAPTIVE MODULATION AND RECEIVER DESIGN FOR CDMA FADING CHANNELS WITH UNCERTAIN CHANNEL STATE INFORMATION 1
R.J. Barton and J. Ni

IMPROVED LINEAR RECEIVERS FOR BPSK-CDMA SUBJECT TO FADING 11
A.M. Tulino and S. Verdú

PERFORMANCE ANALYSIS OF SUBSPACE BASED SYNCHRONIZATION ALGORITHMS FOR MULTI-RATE DS-CDMA SYSTEMS 22
Z. Pi and U. Mitra

PERFORMANCE OF BLIND AND GROUP-BLIND MULTIUSER DETECTORS 32
A. Høst-Madsen and X. Wang

ON THE DIFFERENCE OF TWO CHI-SQUARE VARIATES WITH APPLICATION TO OUTAGE PROBABILITY COMPUTATION 42
M.K. Simon and M.-S. Alouini

LOW-COMPLEXITY BANDWIDTH-EFFICIENT MUI/ISI-RESILIENT CDMA BASED ON BLOCK-SPREADING 52
S. Zhou, G.B. Giannakis, and C. Le Martret

I-B: CODING I: CHANNEL CODING

Organizers: R. Koetter, R.E. Blahut, and N. Boston
(University of Illinois at Urbana-Champaign)
Chair: R. Koetter
(University of Illinois at Urbana-Champaign)

KNOWING WHEN TO STOP 62
B.J. Frey

SOFT-INPUT/SOFT-OUTPUT SEQUENTIAL DECODING 69
G. Lanzetta, H.A. Cabral, and D.J. Costello, Jr.

RECURSIVE CONSTRUCTIONS AND THEIR MAXIMUM LIKELIHOOD DECODING 71
I. Dumer and K. Shabunov

LINEAR-TIME DECODABLE, CAPACITY ACHIEVING BINARY CODES WITH EXPONENTIALLY FALLING ERROR PROBABILITY 81
A. Barg and G. Zémor

I-C: NETWORK OPTIMIZATION
Chair: R. Srikant
(University of Illinois at Urbana-Champaign)

A FUNDAMENTAL NEED FOR DIFFERENTIATED "QUALITY OF SERVICE" OVER COMMUNICATION LINKS: AN INFORMATION THEORETIC APPROACH 83
A. Sahai and S. Mitter

CHANNEL CODING WITH FEEDBACK 93
S. Tatikonda and S. Mitter

ENERGY EFFICIENCY, ENERGY CONSTRAINTS, AND ADMISSION CONTROL IN WIRELESS NETWORKS: A CASE STUDY IN SESSION MULTICASTING 103
J.E. Wieselthier, G.D. Nguyen, and A. Ephremides

ADMISSION CONTROL SCHEMES TO PROVIDE CLASS-LEVEL QOS IN MULTISERVICE NETWORKS 113
S. Kalyanasundaram, E.K.P. Chong, and N.B. Shroff

OPTIMAL ROUTING ALGORITHMS IN AD HOC WIRELESS NETWORKS 123
C.G. Lott and D. Teneketzis

ANALYSIS OF LOW EARTH ORBIT SATELLITE NETWORKS WITH INTER-SATELLITE LINKS 133
A. Sen, R. Jost, and J.M. Capone

I-D: NETWORKED AND OPTIMAL CONTROL
Organizers: C. Beck, G. Dullerud, and P.G. Voulgaris
(University of Illinois at Urbana-Champaign)
Chair: G. Dullerud
(University of Illinois at Urbana-Champaign)

CONTROL OVER NETWORKS 143
H. Ishii and B.A. Francis

FLOW CONTROL VIA PRICING: A FEEDBACK PERSPECTIVE 153
F. Paganini

A NEW APPROACH TO MULTI-OBJECTIVE CONTROLLER SYNTHESIS 163
X. Qi, M.H. Khammash, and M.V. Salapaka

PERFORMANCE EVALUATION OF DISTRIBUTED CONTROL SYSTEMS WITH REDUCED COMMUNICATION 172
J.K. Yook and D.M. Tilbury

CONTROL UNDER COMMUNICATION CONSTRAINTS 182
S. Tatikonda and S. Mitter

I-F: CAPACITY AND DIVERSITY

Chair: A.M. Sayeed
(University of Wisconsin, Madison)

ON BANDWIDTH-EFFICIENT COMMUNICATION OVER FREQUENCY-SELECTIVE FADING CHANNELS 191
A. Ganesan and A.M. Sayeed

ITERATIVE QR DETECTION FOR AN UNCODED SPACE-TIME COMMUNICATION ARCHITECTURE 193
M.O. Damen, K. Abed-Meraim, and S. Burkykh

COMBINED DIRECTIONALITY AND TRANSMIT DIVERSITY VIA SMART ANTENNA SPATIAL SWEEPING 203
S.A. Zekavat, C.R. Nassar, and S. Shattil

SWITCHING BETWEEN MULTIPLEXING AND DIVERSITY BASED ON CONSTELLATION DISTANCE 212
R.W. Heath, Jr. and A. Paulraj

CAPACITY AND ERROR RATE OF SPATIAL CDMA FOR MULTIPLE ANTENNA MULTIPLE ACCESSING 222
J.Y. Hui

FACTORS INFLUENCING THE SHANNON CAPACITY OF SPACE TIME CODE 225
J.Y. Hui, C. Bi, and H. Sun

EFFICIENT USE OF WIDEBAND SPATIO-TEMPORAL WIRELESS CHANNELS GIVEN CHANNEL STATE INFORMATION 228
E.N. Onggosanusi, B.D. Van Veen, and A.M. Sayeed

II-A: CODING II: ITERATIVE DECODING

Organizers: R. Koetter, R.E. Blahut, and N. Boston
(University of Illinois at Urbana-Champaign)
Chair: D. Forney
(Massachusetts Institute of Technology)

CONSTRUCTION TECHNIQUES FOR GRAPHS WITH LARGE GIRTH AND LDPC 238
I.F. Blake

CONSTRUCTIONS OF LDPC CODES USING RAMANUJAN GRAPHS AND IDEAS FROM MARGULIS 248
J. Rosenthal and P.O. Vontobel

IRREGULAR SINGLE PARITY CHECK PRODUCT CODES 258
D. Rankin and T.A. Gulliver

PERFORMANCE ANALYSIS AND CODE OPTIMIZATION OF LOW DENSITY PARITY-CHECK CODES ON RAYLEIGH FADING CHANNELS 266
J. Hou, P.H. Siegel, and L.B. Milstein

THE ASYMPTOTIC INPUT-OUTPUT WEIGHT DISTRIBUTION OF CONVOLUTIONAL ENCODERS 276
I. Sason, E. Telatar, and R. Urbanke

II-B: WIRELESS COMMUNICATIONS II: STYSTEM CONSIDERATIONS

Organizer: V.V. Veeravalli
(University of Illinois at Urbana-Champaign)
Chair: R.J. Barton
(University of Houston)

MAI CONDITIONAL WEAK CONVERGENCE AND NETWORK CAPACITY IN WIRELESS NETWORKS WITH MMSE RECEIVERS 287
J. Zhang and E.K.P. Chong

ADAPTIVE MODULATION TECHNIQUES IN CODED CDMA 297
F. Alesiani, E. Biglieri, G. Taricco, and E. Viterbo

ENSURING CONVERGENCE OF THE MMSE ITERATION FOR INTERFERENCE AVOIDANCE TO THE GLOBAL OPTIMUM 307
P. Anigstein and V. Anantharam

OPTIMUM SIGNATURE SEQUENCE SETS FOR ASYNCHRONOUS CDMA SYSTEMS 317
S. Ulukus and R.D. Yates

RANDOMIZED POWER CONTROL STRATEGIES FOR OPTIMIZATION OF MULTIPLE ACCESS RADIO SYSTEMS 327
G. del Angel and T.L. Fine

ADAPTIVE CODING WITH PILOT SIGNALS 337
M. Médard, I. Abou-Faycal, and U. Madhow

II-C: COMMUNICATION SYSTEMS

Chair: D.L. Goeckel
(University of Massachusetts, Amherst)

OPTIMIZATION OF THE SPREADING BANDWIDTH FOR SELECTIVE RAKE RECEPTION OVER RAYLEIGH FADING CHANNELS 347
W.C. Lau, M.-S. Alouini, and M.K. Simon

PERFORMANCE ANALYSIS OF HIERARCHICAL CODED 16-QAM SYSTEMS 357
J.-H. Lim and S.B. Gelfand

OUTAGE PROBABILITY OF WIRELESS COMMUNICATION SYSTEMS WITH A MINIMUM SIGNAL POWER CONSTRAINT: SOME CLOSED-FORM EXPRESSIONS 367
H. Yang and M.-S. Alouini

A NOVEL HYBRID SINGLE-CARRIER/MULTICARRIER SYSTEM FOR HIGHLY BANDWIDTH EFFICIENT WIRELESS COMMUNICATION 377
G. Ananthaswamy and D.L. Goeckel

PERFORMANCE ANALYSIS OF DISTRIBUTED DYNAMIC CHANNEL ASSIGNMENT WITH POWER CONTROL FOR CELLULAR MOBILE RADIO SYSTEMS 387
Z.H. Zheng and W.H. Lam

OPTIMAL DISTRIBUTED POWER CONTROL IN CELLULAR WIRELESS SYSTEMS 393
L. Qian and Z. Gajic

II-D: CONTROL THEORY
Chair: P.G. Voulgaris
(University of Illinois at Urbana-Champaign)

REACHABILITY UNDER DISTURBANCES 403
A.B. Kurzhanski and P. Varaiya

ON OPTIMAL INPUT DESIGN FOR PARAMETER ESTIMATION PROBLEMS IN PDE'S 412
P.L. Chow and R. Khasminskii

LINEAR CUMULANT CONTROL AND ITS RELATIONSHIP TO RISK-SENSITIVE CONTROL 422
L. Mou, S.R. Liberty, K.D. Pham, and M.K. Sain

REDUCED-ORDER OBSERVER DESIGN FOR UNCERTAIN DISCRETE-TIME STOCHASTIC SYSTEMS WITH GENERAL CRITERIA 431
E.E. Yaz and Y.I. Yaz

A SIMPLIFIED TIME MAXIMUM DISTURBANCE SWITCH CURVE FOR SECOND ORDER SYSTEMS 439
K.H. You and E.B. Lee

IDENTIFYING CONSTRAINED RECEDING HORIZON CONTROLLERS 449
S. Gentry, S. Venkatesh, and E. Feron

STOCHASTIC CONTROL UNDER COMMUNICATION CONSTRAINTS 451
S. Tatikonda and S. Mitter

II-E: SOURCE CODING
Organizers/Chairs: A. Singer and D. Baron
(University of Illinois at Urbana-Champaign)

ON RATE-DISTORTION THEORY AS A CONVEX SELECTION PROBLEM 461
I. Kontoyiannis

NEW RESULTS IN UNIVERSAL CODING OF SIMPLE NON-STATIONARY SOURCE CLASSES 468
G.I. Shamir and D.J. Costello, Jr.

ON UNIVERSAL COMPRESSION OF MULTI-DIMENSIONAL DATA ARRAYS USING SELF-SIMILAR CURVES 470
T. Weissman and S. Mannor

STRUCTURALLY EQUIVALENT GRAMMARS IN LOSSLESS DATA COMPRESSION 480
J. Kieffer and E.-H. Yang

LOSSLESS COMPRESSION OF RICH ALPHABET SOURCES VIA BACKWARD PROCESSING 488
I. Sutskover, B. Oklander, E. Surijan, and J. Ziv

UNIFORMLY IMPROVING THE GENERALIZED LIKELIHOOD RATIO TEST FOR LINEAR GAUSSIAN CHANNELS 498
E. Erez and M. Feder

II-F: NETWORK ANALYSIS AND CONTROL

Chair: P.R. Kumar
(University of Illinois at Urbana-Champaign)

ON OPTIMAL CONTROL OVER PACKET-ORIENTED COMMUNICATION LINKS 508
M. Johansson and L. Xiao

STOCHASTIC TRANSMISSION POLICIES FOR TIME-VARYING CHANNELS WITH MEMORY UNDER INCOMPLETE CHANNEL STATE INFORMATION 518
D. Zhang and K.M. Wasserman

DISCRETE-TIME SINGULARLY PERTURBED MARKOV CHAINS 528
G. Badowski, G. Yin, and Q. Zhang

ON FLUID ANALYSIS OF QUEUES WITH SELECTIVE MESSAGE DISCARDING POLICIES 537
P. Dube and E. Altman

EQUIVALENT MODELS AND ANALYSIS FOR MULTI-STAGE TREE NETWORKS OF DETERMINISTIC SERVICE TIME QUEUES 547
M.J. Neely and C.E. Rohrs

ESTIMATING THE LOSS PROBABILITY FOR FINITE BUFFER MULTIPLEXERS 557
H.S. Kim and N.B. Shroff

III-A: NETWORK MONITORING

Organizer/Chair: C. Hadjicostis
(University of Illinois at Urbana-Champaign)

LOAD ANOMALY DETECTION FOR VOICE OVER IP TRAFFIC 567
M. Mandjes, I. Saniee, and A.L. Stolyar

A GENERIC ALGORITHM FOR MONITORING VARIOUS STOCHASTIC NETWORK CHARACTERISTICS 577
A. Burrell and P. Papantoni-Kazakos

LIMITS OF BER GUARANTEES IN TRANSPARENT OPTICAL NETWORKS USING OOK SIGNALING 583
P. Saengudomlert and M. Médard

III-B: CODING III: ALGEBRAIC DECODING

Organizers: R. Koetter, R.E. Blahut, and N. Boston
(University of Illinois at Urbana-Champaign)
Chair: R.E. Blahut
(University of Illinois at Urbana-Champaign)

A FAST SPECIAL FACTORIZATION ALGORITHM IN THE SUDAN DECODING PROCEDURE..............593
G.-L. Feng

COMBINATORIAL BOUNDS FOR LIST DECODING..............603
V. Guruswami, J. Hastad, M. Sudan, and D. Zuckerman

THE DISPLACEMENT METHOD IN CODING THEORY..............613
V. Olshevsky and A. Shokrollahi

ALGEBRAIC SOFT-DECISION DECODING OF REED-SOLOMON CODES..............625
R. Koetter and A. Vardy

III-C: ESTIMATION

Chair: P. Moulin
(University of Illinois at Urbana-Champaign)

ON THE (NON-)CONVERGENCE OF THE EM ALGORITHM FOR DISCRETE PARAMETER ESTIMATION..............636
P. Spasojević and C.N. Georghiades

AN ALTERNATING MINIMIZATION APPROACH TO IMAGE FORMATION IN THE PRESENCE OF A KNOWN FUNCTION..............646
R.J. Murphy, J.A. O'Sullivan, and D.L. Snyder

SPATIAL SPECTRUM ESTIMATION FOR TIME-VARYING ARRAYS USING THE EM ALGORITHM....648
D.W. Rieken, A.D. Lanterman, and D.R. Fuhrmann

III-D: EQUALIZATION AND SPACE-TIME CODING

Chair: A. Singer
(University of Illinois at Urbana-Champaign)

EQUALIZATION OF SPACE-TIME CODED MULTI-ANTENNA SYSTEMS..............658
H. Viswanathan

A NEW CLASS OF EFFICIENT BLOCK-ITERATIVE INTERFERENCE CANCELLATION TECHNIQUES FOR DIGITAL COMMUNICATION RECEIVERS..............668
A.M. Chan and G.W. Wornell

COMBINED FIR MIMO DECISION FEEDBACK EQUALIZATION AND SPACE-TIME BLOCK CODING OVER MULTIPATH FADING CHANNELS..............680
N. Al-Dhahir and A.F. Naguib

CODES FOR UNKNOWN ISI CHANNELS..............689
O. Coskun and K.M. Chugg

III-F: WIRELESS COMMUNICATIONS III: CHANNEL MODELING, INFORMATION THEORY

Organizer: V.V. Veeravalli
(University of Illinois at Urbana-Champaign)
Chair: A. Goldsmith
(Stanford University)

SCATTERING OF ULTRA-WIDEBAND SIGNALS 700
T. Berger and J. Cheng

ON MODELING MULTI-ANTENNA MULTIPATH CHANNELS 705
A.M. Sayeed

ON ACHIVABLE RATES IN A MULTI-ANTENNA BROADCAST DOWNLINK 715
G. Caire and S. Shamai

ADAPTIVE PARAMETER ESTIMATION IN CDMA SYSTEMS WITH APERIODIC CODES 725
S. Buzzi and H.V. Poor

IV-A: WIRELESS COMMUNICATIONS IV: NETWORKING

Organizer: V.V. Veeravalli
(University of Illinois at Urbana-Champaign)
Chair: E.K.P. Chong
(Purdue University)

MOBILITY INCREASES THE CAPACITY OF AD-HOC WIRELESS NETWORKS 735
M. Grossglauser and D. Tse

SOME CAPACITY RESULTS FOR AD HOC NETWORKS 745
S. Toumpis and A. Goldsmith

COOPERATIVE DETECTION AND COMMUNICATION IN WIRELESS SENSOR NETWORKS 755
M. Ahmed, Y.-S. Tu, and G. Pottie

AN ADAPTIVE MULTI-QUEUE SERVICE ROOM PROTOCOL FOR WIRELESS NETWORKS WITH MULTIPACKET RECEPTION 765
Q. Zhao and L. Tong

EXPLOITING DISTRIBUTED SPATIAL DIVERSITY IN WIRELESS NETWORKS 775
J.N. Laneman and G.W. Wornell

LOW ENERGY WIRELESS COMMUNICATION NETWORK DESIGN 785
W. Stark, H. Wang, A. Worthen, P. Liang, R. Gupta, J. East, A. Hero, S. Lafortune, and D. Teneketzis

IV-B: CDMA

Chair: B. Hajek
(University of Illinois at Urbana-Champaign)

A COMPARISON OF SIGNAL DESIGN METHODS FOR CORRELATED WAVEFORM MULTIPLE-ACCESS COMMUNICATIONS 795
T. Guess and M.K. Varanasi

COMBINED CDMA SIGNAL DESIGN AND POWER CONTROL FOR DECISION-FEEDBACK RECEIVERS SUBJECT TO A QUALITY-OF-SERVICE CONSTRAINT 806
T. Guess

GROUP DETECTORS FOR DS/CDMA SYSTEMS WITH MULTIPATH FADING CHANNELS 816
S. Buzzi, M. Lops, A. Pauciullo, and G. Ricci

A FAMILY OF LIKELIHOOD ASCENT SEARCH DETECTORS ACHIEVING LOCAL MAXIMUM LIKELIHOOD WITH AN ARBITRARY NEIGHBORHOOD SIZE FOR CDMA MULTIUSER DETECTION 826
Y. Sun

MULTIUSER DETECTORS FOR INCOMPLETELY KNOWN SYMMETRIC SIGNALS 836
Y. Li and A. Ephremides

A NOVEL SUBOPTIMAL CDMA MULTIUSER DETECTOR 845
H.F.F. Lu and P.V. Kumar

BLIND ADAPTIVE LINEAR MULTIUSER DETECTION FOR VARIABLE SPREADING FACTOR CDMA 847
N. Seidl, I. Howitt, and J. Richie

IV-C: SIMULATION AND OPTIMIZATION
Organizers/Chair: S.G. Henderson and S.P. Meyn
(University of Michigan and University of Illinois at Urbana-Champaign)

ON THE NUMBER OF SAMPLES REQUIRED FOR Q-LEARNING 849
V.S. Borkar

FIXED POINTS OF APPROXIMATE VALUE ITERATION AND TEMPORAL-DIFFERENCE LEARNING 855
D.P. de Farias and B. Van Roy

SOME APPLICATIONS OF REINFORCEMENT LEARNING TO REAL-WORLD PROBLEMS 865
A. Gosavi

ENHANCED OPTIMALITY CONDITIONS AND EXACT PENALTY FUNCTIONS 872
D.P. Bertsekas and A.E. Koksal

SLOWEST DESCENT SEQUENCE ESTIMATION FOR PROBLEMS WITH A QUADRATIC LOG-LIKELIHOOD 886
P. Spasojević and C.N. Georghiades

IV-D: MOTION PLANNING
Organizer/Chair: F. Bullo
(University of Illinois at Urbana-Champaign)

CONTROLLABILITY AND STABILIZABILITY OF SINGLE INPUT PLANAR JUGGLING 896
K.M. Lynch and C.K. Black

THEORETICAL AND EXPERIMENTAL INVESTIGATION OF STRATIFIED ROBOTIC FINGER GAITING AND MANIPULATION 906
B. Goodwine and Y. Wei

REAL-TIME MOTION PLANNING FOR AUTONOMOUS VEHICLES 916
E. Frazzoli, M.A. Dahleh, and E. Feron

USING RANDOMIZATION TO FIND AND OPTIMIZE FEASIBLE TRAJECTORIES FOR NONLINEAR SYSTEMS 926
P. Cheng, Z. Shen, and S.M. LaValle

MOTION PLANNING FOR A CLASS OF HYBRID LOCOMOTION SYSTEMS 936
M. Žefran

IV-E: TURBO CODES AND ITERATIVE CODING
Chair: R. Koetter
(University of Illinois at Urbana-Champaign)

TWO-DIMENSIONAL INTERLEAVING BY USING SUCCESSIVE PACKING 945
Y.-Q. Shi and X. Zhang

ARRAY CODES AS LOW-DENSITY PARITY-CHECK CODES 955
J.L. Fan

CONVOLUTIONAL CODING USING FINITE-FIELD WAVELETS 957
F. Fekri, S.W. McLaughlin, R.M. Mersereau, and R.W. Schafer

LOW-LATENCY SISO VIA MESSAGE PASSING ON A BINARY TREE 959
P. Thiennviboon and K.M. Chugg

LOW RATE PRODUCT CODES 961
T. Ji and W.E. Stark

AN IMPROVED ITERATIVE DECODING ALGORITHM FOR TURBO PRODUCT CODE 963
X. Wei and A.N. Akansu

ENCODER FOR ROTATIONALLY-INVARIANT TURBO TCM 965
J. Sun, O.Y. Takeshita, and M.P. Fitz

IV-F: SPACE-TIME CODES
Chair: D. Sarwate
(University of Illinois at Urbana-Champaign)

EFFICIENT IMPLEMENTATION OF A MAXIMUM-LIKELIHOOD DETECTOR FOR SPACE-TIME CODED SYSTEMS 967
M. Uysal and C.N. Georghiades

PERFORMANCE ANALYSIS OF BICM-ID WITH SIGNAL SPACE DIVERSITY IN RAYLEIGH FADING 975
A. Chindapol and J.A. Ritcey

SPACE-TIME CODING FOR RAYLEIGH FADING CHANNELS IN CPM SYSTEM 985
X. Zhang and M. Fitz

SPACE-TIME CODING FOR FREQUENCY SELECTIVE FADING CHANNEL 995
Y. Liu, M.P. Fitz, and O.Y. Takeshita

PERFORMANCE OF SIMPLE SPACE-TIME BLOCK CODES FOR MORE THAN TWO TRANSMIT ANTENNAS 997
B. Raghothaman, A. Boariu, O. Tirkkonen, and A. Hottinen

AN INTERFERENCE AND FADING RESILIENT SPACE-TIME CODED CDMA SCHEME 1000
H. Li and J. Li

V-A: CODING IV: ALGEBRAIC AND COMBINATORIAL CODING THEORY

Chair: R.S. Sreenivas
(University of Illinois at Urbana-Champaign)

DESIGNS IN ADDITIVE CODES OVER GF(4) 1010
V. Pless

A CRITICAL LOOK AT SELF-DUAL CODES 1019
J.L. Walker

THE DERIVATIVES OF BOOLEAN FUNCTIONS: SOME RELATED PROBLEMS IN CODING THEORY AND CRYPTOGRAPHY 1029
P. Charpin

A LOW COMPLEXITY ALGORITHM FOR THE CONSTRUCTINON OF ALGEBRAIC GEOMETRIC CODES BETTER THAN THE GILBERT-VARSHAMOV BOUND 1031
K. Shum, I. Aleshnikov, P.V. Kumar, and H. Stichtenoth

UNDERSTANDING LINEAR CODES OF CONSTANT WEIGHT USING VIRTUAL LINEAR CODES 1038
J.A. Wood

V-B: WIRELESS COMMUNICATIONS V: SPACE-TIME COMMUNICATIONS

Organizer: V.V. Veeravalli
(University of Illinois at Urbana-Champaign)
Chair: X. Wang
(Texas A&M University)

HIGH-RATE LINEAR SPACE-TIME CODES 1047
B. Hassibi and B.M. Hochwald

ROBUST SPACE-TIME CODING FOR CORRELATED RAYLEIGH FADING CHANNELS 1057
S. Siwamogsatham and M.P. Fitz

SIGNAL CONSTELLATIONS FOR NONCOHERENT SPACE-TIME COMMUNICATIONS 1067
M.L. McCloud, M. Brehler, and M.K. Varanasi

SPACE-TIME CODED GENERALIZED MULTICARRIER CDMA WITH BLOCK-SPREADING FOR MULTIRATE SERVICES 1076
A. Stamoulis, Z. Liu, and G.B. Giannakis

UNION BOUND FOR LINEAR SPACE-TIME CODES 1086
S. Sandhu, R.W. Heath, Jr., and A. Paulraj

A RANDOM MATRIX MODEL OF COMMUNICATION THROUGH ANTENNA ARRAYS 1096
R.R. Müller

V-C.a: MARKOV PROCESSES AND SIMULATION
Organizers/Chairs: S.G. Henderson and S.P. Meyn
(University of Michigan and University of Illinois at Urbana-Champaign)

BIOMOLECULAR CONFORMATIONS AS METASTABLE SETS OF MARKOV CHAINS 1106
C. Schütte and W. Huisinga

PERFECT SAMPLING AND QUEUEING MODELS 1116
R.L. Tweedie and J.N. Corcoran

PERFECT SIMULATION FOR MARKOV CHAINS ARISING FROM DISCRETE-EVENT SIMULATION 1125
S.G. Henderson and R.L. Tweedie

V-C.b: HYBRID/DISCRETE EVENT SYSTEMS
Chair: R.S. Sreenivas
(University of Illinois at Urbana-Champaign)

FAULT DIAGNOSIS AND CONSISTENCY IN HYBRID SYSTEMS 1135
S.H. Zad, R.H. Kwong, and W.M. Wonham

INVARIANCE UNDER SCALING OF TIME BOUNDS IN TIMED DISCRETE-EVENT SYSTEMS 1145
S.E. Bourdon, W.M. Wonham, and M. Lawford

MODELLING OF BATCH PROCESS PLANTS BY FSMS 1155
S. Knap, T. Shepard, and M.M. Bayoumi

MODIFICATION TO THE BRANDIN-WONHAM TIMED MODEL OF DESS 1165
S. Knap, T. Shepard, and M.M. Bayoumi

V-D: ADAPTIVE ALGORITHMS
Organizer/Chair: A. Singer
(University of Illinois at Urbana-Champaign)

AN ADAPTIVE FILTER ROBUST TO DATA UNCERTAINTIES 1175
T.Y. Al-Naffouri and A.H. Sayed

SOME ASPECTS OF MULTISENSOR FUSION OVER FADING CHANNELS 1184
H. Papadopoulos

A STOCHASTIC ALGORITHM FOR THE JOINT OPTIMIZATION OF POWER CONTROL AND MULTIUSER RECEIVERS 1194
D. Das and M.K. Varanasi

EXISTENCE AND PERFORMANCE OF SHALVI-WEINSTEIN ESTIMATORS 1204
P. Schniter

TIME-REVERSAL DIVERSITY IN DECISION FEEDBACK EQUALIZATION 1214
J. Balakrishnan and C.R. Johnson, Jr.

APPLICATION OF REDUCED-RANK CHIP-LEVEL MMSE EQUALIZATION TO THE FORWARD LINK DS-CDMA WITH FREQUENCY SELECTIVE MULTIPATH 1224
S. Chowdhury, M.D. Zoltowski, and J.S. Goldstein

“TURBO EQUALIZATION” USING FREQUENCY DOMAIN EQUALIZERS 1234
M. Tüchler and J. Hagenauer

V-E: MULTIMEDIA SIGNAL PROCESSING

Organizer/Chair: T.S. Huang
(University of Illinois at Urbana-Champaign)

MULTIMODAL HUMAN-COMPUTER INTERFACES 1244
J.L. Flanagan

SYNTHETIC TALKING FACE DRIVEN BY MULTIMODAL SIGNAL 1247
S. Morishima

MULTIMODAL INFORMATION FUSION FOR IMPROVED SPEAKER DETECTION 1256
A. Garg, V. Pavlović, and J.M. Rehg

V-F: OPTICAL NETWORKS

Organizer/Chair: M. Médard
(Massachusetts Institute of Technology)

SIMPLE LAYOUT ALGORITHMS TO MAINTAIN NETWORK CONNECTIVITY UNDER FAULTS 1266
G.H. Sasaki, C.-F. Su, and D. Blight

ON TRADING WAVELENGTHS WITH FIBERS: A COST-PERFORMANCE BASED STUDY 1274
A.K. Somani and M. Mina

TOPOLOGY DESIGN IN WDM RINGS TO MINIMIZE ELECTRONIC ROUTING: EFFICIENT COMPUTATION OF TIGHT BOUNDS 1284
R. Dutta and G.N. Rouskas

MULTI-COST ROUTING IN MAX-MIN FAIR SHARE NETWORKS 1294
F.J. Gutierrez, E.A. Varvarigos, and S. Vassiliadis

ON ROUTING AND WAVELENGTH ASSIGNMENT IN OPTICAL NETWORKS 1305
A.W. Richa, A. Sen, B.H. Shen, and S. Bandyopadhyay

ON THE EFFECTS OF SWITCHING IN MULTI-FIBER WDM NETWORKS 1314
S. Subramaniam, H. Choi, and H.-A. Choi

LIST OF AUTHORS 1324

Adaptive Modulation and Receiver Design for CDMA Fading Channels with Uncertain Channel State Information

Richard J. Barton and Jianjun Ni
{rbarton@uh.edu} and {jjni@iastate.edu}
Electrical and Computer Engineering Department
University of Houston
Houston, TX 77204

Abstract---In this paper, the design and analysis of adaptive linear-quadratic (LQ) receivers for time-varying, frequency-selective CDMA environments with a high level of uncertainty are studied. We propose a simple, intuitively appealing cost function that can be maximized to find LQ detectors that are optimal in a certain sense. We discuss the motivation and derivation of the proposed LQ cost function and derive an adaptive algorithm for LQ detectors that can be viewed as a generalization of the minimum-output-energy algorithm for the MMSE linear detector. An approach to adaptive signal design for time-varying channels based on the proposed LQ cost function is also discussed. Finally, we present the results of a simulation study that clearly demonstrates the potential performance advantage of the proposed LQ receivers vis-à-vis MMSE linear receivers.

I. INTRODUCTION

Third-generation mobile radio systems (*e.g.*, IMT-2000/UMTS), are envisioned as a means of providing ubiquitous mobile personal multimedia services in a convergent environment in which the consumer will no longer choose between delivery networks but between the services that meet his needs. Such systems represent one of the key communication technologies of interest to research, development, and international standardization bodies due to the high market demand for advanced wireless communications. These systems will operate in low-power, time-varying, multiuser environments in which the exact structure of the channel is difficult to determine precisely. We refer to this environment generically as the *mobile wireless channel*, and it serves as a primary motivation for this research.

In this paper, we discuss the design of adaptive receivers for direct-sequence, code-division multiple-access (DS-CDMA or just CDMA) channels. The work is motivated by the utility and popularity of CDMA as a means for implementation of an all-digital wireless system. Due to the complex and rapidly time-varying structure of the mobile wireless environment, it is often difficult or impossible to estimate channel characteristics accurately. Hence, it is often necessary to implement a receiver in a heavily cluttered environment characterized by a high level of channel uncertainty. We consider a novel Bayesian approach to this problem in which adaptive linear-quadratic (LQ) receivers are designed based on a probabilistic channel model that explicitly incorporates a stochastic structure for the channel uncertainty.

Since we are primarily interested in low-complexity decision strategies that can be implemented in small, low-power receivers, we will restrict our discussion to one-shot, single-user receivers; however, the techniques that we propose to study can also be applied to joint multi-user detection as well as to block detection of multiple bits at the cost of increased complexity. Further, although the ideas developed here can be extended straightforwardly and beneficially to differentially encoded CDMA waveforms, we will confine our discussion to non-differentially encoded systems.

To enhance the performance of the LQ receivers, we investigate adaptive modulation design in parallel with the adaptive receiver design. In this paper, we use the term *adaptive* *modulation* to refer to techniques that alter the size or shape of the signal constellation in response to characteristics of the transmission environment. In particular, we adopt M-ary quadrature amplitude modulation (MQAM) as the basic signaling format and consider techniques to determine adaptively the number M of symbols in the constellation and the coordinates of each symbol with respect to a known set of basis signals. Note that, in principle, our definition of adaptive modulation encompasses adaptive power control as well since the average transmitted power of the signaling scheme is one of the parameters determined by the coordinates of the signal constellation. Also note that we are primarily concerned at this point with uncoded modulation schemes.

Typically, the goal of an adaptive modulation scheme on a communication channel is related to one of the following:

- Maximize the throughput on the channel subject to constraints on the transmitted power and the BER.
- Minimize the transmitted power subject to constraints on the throughput and BER.
- Minimize the BER subject to constraints on the throughput and transmitted power.

Adaptive modulation techniques addressing variations of these problems have been studied by many different investigators. Of particular interest here are techniques designed to combat the effects of channel fading, such as those studied in [1-7]. For example, in [7], the signal constellation is chosen from a discrete set of possibilities with fixed average power in order to maintain a constant BER with a high data rate (large constellation size) when the channel is favorable (shallow fade) and a low data rate (small constellation size) when the channel is unfavorable (deep fade). Similarly, in [2], both the constellation size and the transmitted power are varied to maintain constant BER and constant average transmitted power but maximize the data rate for any particular channel state.

The paper is organized as follows. The assumed Bayesian channel model is introduced in Section 2. In Section 3, we motivate and derive the proposed LQ receivers. Simulation results demonstrating the performance of the LQ receivers vis-à-vis MMSE linear receivers are presented in Section 4. Finally, in Section 5, we present some concluding remarks and discuss some interesting aspects that are currently under investigation.

II. SYSTEM MODEL

We assume initially that information is transmitted using binary (but not necessarily antipodal) direct-sequence CDMA modulation. Since we will be concerned primarily with noncoherent systems, we adopt a complex-valued baseband model for the signal. That is, we assume that the transmitted baseband signal for the user of interest (user 0) takes the form

$$s(t)=\sqrt{2P}\sum_{k=-\infty}^{\infty}\sum_{i=0}^{N-1}\left[b_k c_{1,i}+(1-b_k)c_{0,i}\right]\psi(t-\tau-kT-iT_c)e^{i\theta}, \qquad (1)$$

where P is the average power of user 0, N is the length of the chip sequence, $T_c=T/N$ is the length of the chip interval, $b_k\in\{0,1\}$ represents the transmitted value of bit k for user 0, $\psi(t)$ is a chip waveform of duration T_c, τ is the relative delay for user 0, θ is the relative phase of user 0, and $c_{j,i}$ is the ith component of the user's chip sequence $\mathbf{c}_j$ corresponding to a transmitted bit value of $j\in\{0,1\}$. Without loss of generality, we assume that $\theta=0$ and $\tau=0$. We also assume that the chip sequences $\mathbf{c}_0$ and $\mathbf{c}_1$ are known to the receiver. Note that the choice of $\mathbf{c}_0=-\mathbf{c}_1$ corresponds to an antipodal binary signaling structure, while choosing $\mathbf{c}_0$ and $\mathbf{c}_1$ to be orthogonal spreading sequences corresponds to an orthogonal signaling structure.

To highlight the most salient features of the proposed research, we will confine our discussion to a simple discrete-time model for the channel. This is equivalent, for example, to the assumption that the receiver demodulates the received baseband signal *r*(*t*) into a sequence of symbols $\{r_m\}_{-\infty}^{\infty}$ using a noncoherent chip-matched filter. The received symbol vector corresponding to a single detection interval can then be modeled as:

$$r_m = \sum_{l=0}^{L} \hat{\alpha}_l s_{m-l} + \sum_{l=0}^{L} \varepsilon_l s_{m-l} + \gamma_m, \tag{2}$$

where $\{s_m\}$ represents the sequence of transmitted symbols for user 0, $\{\gamma_m\}$ represents the cumulative effects of all of the additive interference – *i.e.*, multiple-access interference (MAI), narrowband interference (NBI), and additive white Gaussian noise (AWGN) – the sequence $\{\hat{\alpha}_l\}_{l=0}^{L}$ represents *known* estimated values of the discrete-time channel impulse-response sequence for user 0, and the sequence $\{\varepsilon_l\}_{l=0}^{L}$ represents the *unknown* errors in the estimates of the impulse-response sequence.

Now, because we are interested in one-shot detectors, we consider an observation vector $\mathbf{x}(n)$ of length *N*+*L* (for example) with components $x_i(n) = r_{nN+i}$, for $0 \le i \le N+L-1$. Reverting entirely to vector notation and dropping the explicit dependence on the bit interval *n*, we see that the observation vector $\mathbf{x}$ can be written in the form

$$\mathbf{x} = \mathbf{S}\hat{\boldsymbol{\alpha}} + \mathbf{S}\boldsymbol{\varepsilon} + \boldsymbol{\iota} + \boldsymbol{\gamma}, \tag{3}$$

where $\mathbf{S}$ is a matrix of known form determined by $\mathbf{c}_0$, $\mathbf{c}_1$, and the current transmitted bit value for user 0, and $\hat{\boldsymbol{\alpha}}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\iota}$ are vectors representing estimated channel impulse response, channel estimation errors, additive interference, and intersymbol interference (ISI), respectively. To complete the channel model, we make the simplifying (and generally reasonable) assumptions that $\hat{\boldsymbol{\alpha}}$, $\boldsymbol{\varepsilon}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\iota}$ are mutually uncorrelated, wide-sense stationary random vectors, and that $\boldsymbol{\varepsilon}$ has mean zero.

III. ADAPTIVE LQ RECEIVERS

Before deriving the LQ receiver structure, we make one final simplification in the underlying channel model. Since the observation vector $\mathbf{x}$ and the estimated CSI vector $\hat{\boldsymbol{\alpha}}$ are both known to the receiver, it is a simple matter to compute estimates $\boldsymbol{\mu}_{\mathbf{x}}$ and $\boldsymbol{\mu}_{\hat{\alpha}}$ of the corresponding mean vectors in an adaptive fashion. Assuming these estimates are unbiased, we can transform the observation vector $\mathbf{x}$ into a zero-mean observation vector $\mathbf{y}$ using the transformation

$$\mathbf{y} = \mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}} - \tfrac{1}{2}(\mathbf{C}_1 + \mathbf{C}_0)(\hat{\boldsymbol{\alpha}} - \boldsymbol{\mu}_{\hat{\alpha}}), \tag{4}$$

where the $(N+L)\times(L+1)$ matrices $\mathbf{C}_0$ and $\mathbf{C}_1$ are defined straightforwardly in terms of the known spreading sequences $\mathbf{c}_0$ and $\mathbf{c}_1$. The new observation vector then takes the form $\mathbf{y} = \boldsymbol{\mu} b + \boldsymbol{\eta}$, where $b \in \{-1, +1\}$ now represents the transmitted bit value, $\boldsymbol{\mu} = \tfrac{1}{2}(\mathbf{C}_1 - \mathbf{C}_0)\hat{\boldsymbol{\alpha}}$ represents a constant transmitted baseband "signal", and

$$\boldsymbol{\eta} = \tfrac{1}{2} b (\mathbf{C}_1 - \mathbf{C}_0)\boldsymbol{\varepsilon} + \tfrac{1}{2}(\mathbf{C}_1 + \mathbf{C}_0)\boldsymbol{\varepsilon} + \tilde{\boldsymbol{\iota}} + \tilde{\boldsymbol{\gamma}} \tag{5}$$

represents zero-mean additive channel noise. Note, that the ISI and additive interference vectors $\boldsymbol{\iota}$ and $\boldsymbol{\gamma}$ have been replaced with their zero-mean counterparts $\tilde{\boldsymbol{\iota}}$ and $\tilde{\boldsymbol{\gamma}}$. Conceptually then, we have transformed the original detection problem into a more conventional antipodal binary detection problem, which is convenient for expository purposes. Furthermore, given the transmitted bit value b and the estimated CSI vector $\hat{\boldsymbol{\alpha}}$, the observation vector $\mathbf{y}$ has conditional mean vector

$$\boldsymbol{\mu}_{\mathbf{y}|b,\hat{\boldsymbol{\alpha}}} = \boldsymbol{\mu} b = \tfrac{1}{2} b\left(\mathbf{C}_1 - \mathbf{C}_0\right)\hat{\boldsymbol{\alpha}} \tag{6}$$

and conditional covariance matrix

$$\boldsymbol{\Sigma}_{\mathbf{y}|b,\hat{\boldsymbol{\alpha}}} = \begin{cases} \boldsymbol{\Sigma}_0 = \mathbf{C}_0\boldsymbol{\Sigma}_\varepsilon\mathbf{C}_0^* + \boldsymbol{\Sigma}_\iota + \boldsymbol{\Sigma}_\gamma, & b=-1, \\ \boldsymbol{\Sigma}_1 = \mathbf{C}_1\boldsymbol{\Sigma}_\varepsilon\mathbf{C}_1^* + \boldsymbol{\Sigma}_\iota + \boldsymbol{\Sigma}_\gamma, & b=+1, \end{cases} \tag{7}$$

where $\boldsymbol{\Sigma}_\varepsilon$ represents the covariance matrix of the channel uncertainty vector $\boldsymbol{\varepsilon}$, $\boldsymbol{\Sigma}_\gamma$ represents the covariance matrix of the additive interference vector $\boldsymbol{\gamma}$, and $\boldsymbol{\Sigma}_\iota$ represents the covariance matrix of the ISI vector $\boldsymbol{\iota}$. Note that the matrix $\boldsymbol{\Sigma}_\iota$ also has a known form that is completely determined by $\mathbf{c}_0$, $\mathbf{c}_1$, $\boldsymbol{\Sigma}_\varepsilon$, and $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}}$, the covariance matrix of the observed estimated CSI vector $\hat{\boldsymbol{\alpha}}$. However, the corresponding formula is somewhat lengthy and will not be repeated here.

Equations (6) and (7) reveal several interesting properties related to the second-order statistics of the observation vector $\mathbf{y}$. First, in the absence of instantaneous CSI estimates ($\hat{\boldsymbol{\alpha}} = \mathbf{0}$, *i.e.*, completely noncoherent detection), the conditional mean vectors under the binary hypotheses $H_0 : b = -1$ and $H_1 : b = +1$ are identical. Hence, the only second-order information that can be exploited by the receiver to discriminate between the two hypotheses is the difference in the conditional covariance matrices. On the other hand, if some CSI is available ($\hat{\boldsymbol{\alpha}} \neq \mathbf{0}$, *i.e.*, partially or completely coherent detection), then the conditional mean vectors will differ as long as the signal structure under the two hypotheses is not identical ($\mathbf{c}_1 \neq \mathbf{c}_0$). The availability of CSI can obviously be exploited by the receiver by including a linear component in the detector, and the processing gain associated with the linear detector component will be maximized if the signals are antipodal ($\mathbf{c}_1 = -\mathbf{c}_0$).

Similarly, the conditional covariance matrices of $\mathbf{y}$ under the two hypotheses H_0 and H_1 differ only due to the influence of the terms $\mathbf{C}_j\boldsymbol{\Sigma}_\varepsilon\mathbf{C}_j$, $j = 0,1$. This implies that if the CSI is known precisely ($\boldsymbol{\varepsilon} = \mathbf{0}$, *i.e.*, completely coherent detection), then the covariance matrices under the two hypotheses will be identical and the only second-order information that can be exploited by the receiver to discriminate between the two hypotheses is the difference in the mean vectors. However, if the CSI is not known precisely ($\boldsymbol{\varepsilon} \neq \mathbf{0}$, *i.e.*, partially coherent or noncoherent detection) the difference in the covariance matrices under the two hypotheses will depend on the interaction between the signal structure ($\mathbf{c}_0$ and $\mathbf{c}_1$) and the covariance structure of the channel uncertainty vector $\boldsymbol{\varepsilon}$. Hence, the presence of channel uncertainty can actually be exploited by the receiver by including a quadratic component in the detector, but only if the signal structure is chosen appropriately. In particular, if an antipodal signal structure is chosen, channel uncertainty cannot be exploited by the receiver (at least based on second-order statistics) and will serve only to degrade system performance.

In order to exploit both the available CSI and the channel uncertainty in the receiver, we propose to develop adaptive LQ receivers in which the detector takes the form

$$\delta(\mathbf{y}) = \begin{cases} 1, & T(\mathbf{y}) \geq \tau, \\ 0, & T(\mathbf{y}) < \tau, \end{cases} \qquad T(\mathbf{y}) = \mathbf{y}^*\boldsymbol{\Phi}\mathbf{y} + 2\operatorname{Re}\left(\mathbf{h}^*\mathbf{y}\right), \tag{8}$$

where $\boldsymbol{\Phi}$ is a Hermitian (conjugate symmetric) matrix that determines the quadratic component of the detector, $\mathbf{h}$ is a complex-valued vector that determines the linear component of the detector, and τ is a real-valued detection threshold. The first question that must be answered in order to implement such a detector is how to choose the components

$(\boldsymbol{\Phi}, \mathbf{h}, \tau)$. One possible approach to this problem is to choose $(\boldsymbol{\Phi}, \mathbf{h}, \tau)$ to minimize the quantity

$$E\left\{\left[b-\left(T(\mathbf{y})-\tau\right)\right]^2\right\}. \tag{9}$$

This leads to a minimum-mean-squared-error (MMSE) LQ receiver. Note that if an antipodal signal structure is employed, the resulting detector will be equivalent to the more familiar MMSE linear detector [8,9]. Unfortunately, in general, identifying the MMSE LQ detector requires knowledge of the third- and fourth-order moment structure under both hypotheses in addition to the first- and second-order information that is required to identify the linear MMSE detector. Hence, this approach to LQ receiver design is somewhat impractical, particularly if the statistics of the problem are time-varying and must be acquired adaptively.

As an alternative to the MMSE approach, we consider the simple expedient of identifying the optimal (necessarily LQ) detector corresponding to the hypotheses H_0 and H_1 under the additional assumption that the additive noise vector $\boldsymbol{\eta}$ is Gaussian. While this is clearly not an accurate assumption in most cases, it is nevertheless an assumption that often leads to suboptimal receivers with excellent performance characteristics. In particular, if an antipodal signal structure is employed, this approach will again lead to a detector that is equivalent to the MMSE linear detector. Since the MMSE receiver is known to be a robust receiver in the presence of heavy MAI, it can be argued that the assumption of Gaussian additive interference is a good starting point for designing robust detectors even on multiple-access channels. Similarly, it has been shown in [10] that the maximum divergence between an appropriate Gaussian distribution and the MAI-plus-noise component at the output of the linear stage of an MMSE detector is quite small in most cases.

Assuming the matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are invertible, this approach leads to an LQ detector with components given by

$$\begin{aligned}\mathbf{h} &= \left(\boldsymbol{\Sigma}_0^{-1}+\boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{\mu},\\ \boldsymbol{\Phi} &= \boldsymbol{\Sigma}_0^{-1}-\boldsymbol{\Sigma}_1^{-1},\\ \tau &= \ln\left(|\boldsymbol{\Sigma}_1|/|\boldsymbol{\Sigma}_0|\right)-\boldsymbol{\mu}^*\boldsymbol{\Phi}\boldsymbol{\mu}.\end{aligned} \tag{10}$$

While Equations (10) provide a possible solution to the problem of identifying the components of the LQ receiver, it is not a completely satisfying solution for two reasons. First, this solution requires the inversion of the matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. Since the system is time-varying, the matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ must be tracked adaptively, and the inverses must be recomputed accordingly. Since matrix inversion is a computationally intensive operation, frequent recomputation of these inverses is undesirable. While it is possible to track both the matrices and their inverses directly using subspace decomposition techniques (as discussed, for example in [11,12]), it is of interest to identify adaptive solutions for $(\boldsymbol{\Phi}, \mathbf{h}, \tau)$ that do not require inversion of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. Second, the solution given by Equations (10) requires prior knowledge of the signal structure for the problem and does not provide any insight or methodology for choosing a signal structure adaptively in order to optimize receiver performance.

To address both of these deficiencies of solution (10), we consider an alternative approach to adaptive LQ receiver design, which allows us to identify simultaneously both a signal structure $(\mathbf{c}_0, \mathbf{c}_1)$ and a detector structure $(\boldsymbol{\Phi}, \mathbf{h}, \tau)$ that are jointly optimal in a certain sense. Toward this end, we assume for the moment that the matrices $\boldsymbol{\Sigma}_\varepsilon$, and $\boldsymbol{\Sigma}_\gamma$ are known to the receiver, along with the current estimated CSI vector $\hat{\boldsymbol{\alpha}}$ (from which we can also estimate $\boldsymbol{\Sigma}_{\hat{\mathbf{a}}}$). In this case, both the mean vector $\boldsymbol{\mu}$ and the two covariance matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are explicit functions of $\mathbf{c}_0$ and $\mathbf{c}_1$ as given in Equations (6) and (7) above. To simultaneously adapt both the signal structure and the detector structure, we seek a solution $\left(\tilde{\mathbf{c}}_0, \tilde{\mathbf{c}}_1, \tilde{\mathbf{h}}, \tilde{\boldsymbol{\Phi}}\right)$ to the constrained maximization problem

$$\max_{(\mathbf{c}_0, \mathbf{c}_1, \mathbf{h}, \boldsymbol{\Phi})} D(\mathbf{c}_0, \mathbf{c}_1, \mathbf{h}, \boldsymbol{\Phi}) \text{ subject to: } \|\mathbf{c}_0\|^2 = \|\mathbf{c}_1\|^2 = E_b, \tag{11}$$

where $\|\mathbf{c}_j\|^2 = \sum_{i=0}^{N-1} c_{ji}^2$, E_b represents the energy transmitted per bit, and

$$D(\mathbf{c}_0, \mathbf{c}_1, \mathbf{h}, \boldsymbol{\Phi}) = \frac{\left|\boldsymbol{\mu}^*(\mathbf{h}_0 + \mathbf{h}_1) + \frac{1}{4}\mathrm{Tr}\left[(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\boldsymbol{\Phi}\right]\right|^2}{\mathbf{h}_0^*\boldsymbol{\Sigma}_0\mathbf{h}_0 + \mathbf{h}_1^*\boldsymbol{\Sigma}_1\mathbf{h}_1 + \frac{1}{4}\mathrm{Tr}\left[\boldsymbol{\Phi}\boldsymbol{\Sigma}_0\boldsymbol{\Phi}\boldsymbol{\Sigma}_1\right]}, \quad \mathbf{h} = \mathbf{h}_0 + \mathbf{h}_1. \tag{12}$$

The optimal detector structure is completed by defining the threshold $\tilde{\tau}$ as before; that is,

$$\tilde{\tau} = \ln\left(\left|\tilde{\boldsymbol{\Sigma}}_1\right| / \left|\tilde{\boldsymbol{\Sigma}}_0\right|\right) - \tilde{\boldsymbol{\mu}}^*\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\mu}}, \tag{13}$$

We refer to the cost function $D(\mathbf{c}_0, \mathbf{c}_1, \mathbf{h}, \boldsymbol{\Phi})$ as the *modified deflection ratio*. A more restricted version of this cost function has been studied previously in [13], where it was shown to have some desirable properties. For the particular problem of detecting a zero-mean Gaussian signal in the presence of zero-mean Gaussian noise, this cost function is equivalent to the modified deflection ratio discussed in [14,15].

To see that this solution is indeed optimal in some sense, we note that, for a fixed signal structure, the maximum value of the modified deflection ratio is given by

$$D^*(\mathbf{c}_0, \mathbf{c}_1) = \max_{(\mathbf{h}_0, \mathbf{h}_1, \boldsymbol{\Phi})} D(\mathbf{c}_0, \mathbf{c}_1, \mathbf{h}, \boldsymbol{\Phi}) = \boldsymbol{\mu}^*\left(\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_1^{-1}\right)\boldsymbol{\mu} + \tfrac{1}{4}\mathrm{Tr}\left(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1} - 2\mathbf{I}\right), \tag{14}$$

which is attained (except for an arbitrary scaling factor) if and only if $\boldsymbol{\Phi} = \boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Sigma}_1^{-1}$, $\mathbf{h}_0 = \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}$, and $\mathbf{h}_1 = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}$. Hence, for a fixed signal structure, maximizing the modified deflection ratio leads to a detector structure that is equivalent to the optimal Gaussian detector.

This approach also leads to efficient adaptive algorithms. To see this, we consider first the problem of maximizing $D(\mathbf{c}_0, \mathbf{c}_1, \mathbf{h}, \boldsymbol{\Phi})$ for a fixed signal pair $(\mathbf{c}_0, \mathbf{c}_1)$. While an adaptive algorithm could be derived by attempting to maximize the modified deflection ratio directly over all possible sets $(\mathbf{h}_0, \mathbf{h}_1, \boldsymbol{\Phi})$, an easier and more stable approach is based on the fact that the set $\left(\hat{\mathbf{h}}_0, \hat{\mathbf{h}}_1, \hat{\boldsymbol{\Phi}}\right)$ maximizes the modified deflection ratio if and only if (modulo a scaling factor)

$$\hat{\mathbf{h}}_0 = \frac{\tilde{\mathbf{h}}_0}{\tilde{\mathbf{h}}_0^*\boldsymbol{\Sigma}_0\tilde{\mathbf{h}}_0}, \quad \hat{\mathbf{h}}_1 = \frac{\tilde{\mathbf{h}}_1}{\tilde{\mathbf{h}}_1^*\boldsymbol{\Sigma}_1\tilde{\mathbf{h}}_1}, \quad \hat{\boldsymbol{\Phi}} = \frac{\tilde{\boldsymbol{\Phi}}}{\mathrm{Tr}\left[\tilde{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_0\tilde{\boldsymbol{\Phi}}\boldsymbol{\Sigma}_1\right]}, \tag{15}$$

where $\left(\tilde{\mathbf{h}}_0, \tilde{\mathbf{h}}_1, \tilde{\boldsymbol{\Phi}}\right)$ is chosen to solve the dual minimization problem:

$$\begin{aligned} &\min_{\mathbf{h}_0} \mathbf{h}_0^*\boldsymbol{\Sigma}_0\mathbf{h}_0 \text{ subject to } \boldsymbol{\mu}^*\mathbf{h}_0 = 1, \\ &\min_{\mathbf{h}_1} \mathbf{h}_1^*\boldsymbol{\Sigma}_1\mathbf{h}_1 \text{ subject to } \boldsymbol{\mu}^*\mathbf{h}_1 = 1, \\ &\min_{\boldsymbol{\Phi}} \mathrm{Tr}\left[\boldsymbol{\Phi}\boldsymbol{\Sigma}_0\boldsymbol{\Phi}\boldsymbol{\Sigma}_1\right] \text{ subject to } \mathrm{Tr}\left[(\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0)\boldsymbol{\Phi}\right] = 1. \end{aligned} \tag{16}$$

An adaptive solution to this set of constrained minimization problems, which does not involve matrix inversion, can be derived straightforwardly using either the least-mean-square (LMS) algorithm or the recursive least squares (RLS) algorithm. The LMS solution, for example, leads to an adaptive algorithm analogous to the well known minimum-output-energy (MOE) implementation of the linear MMSE detector proposed in [16].

It is straightforward to show that the quantity $D^*(\mathbf{c}_0,\mathbf{c}_1)$ is equivalent to the *divergence distance* between the two Gaussian distributions corresponding to the signal pair $(\mathbf{c}_0,\mathbf{c}_1)$. Hence, solving Problem (11) leads to the optimal Gaussian detector for the signal pair that gives maximum divergence distance between the two corresponding Gaussian hypotheses. Since maximizing the divergence distance is known to be a useful criterion for signal selection in Gaussian detection problems [17-19], this is a desirable result.

To solve simultaneously for the optimal signal structure and the corresponding optimal detector, we note that the value of $D^*(\mathbf{c}_0,\mathbf{c}_1)$ can be rewritten in the form

$$D^*(\mathbf{c}_0,\mathbf{c}_1)=\boldsymbol{\mu}^*(\mathbf{h}_0+\mathbf{h}_1)+\tfrac{1}{4}\mathrm{Tr}(\boldsymbol{\Sigma}_0\boldsymbol{\Phi}\boldsymbol{\Sigma}_1\boldsymbol{\Phi}), \tag{17}$$

where $(\mathbf{h}_0,\mathbf{h}_1,\boldsymbol{\Phi})$ represents the optimal detector for $(\mathbf{c}_0,\mathbf{c}_1)$. Further, assuming that the set $(\mathbf{h}_0,\mathbf{h}_1,\boldsymbol{\Phi})$ is given, the formula for the gradient of $D^*(\mathbf{c}_0,\mathbf{c}_1)$ is straightforward. Hence, to solve simultaneously for the signal structure and the detector structure, we can use a recursive LMS procedure that alternates between updates of $(\mathbf{c}_0,\mathbf{c}_1)$ to solve (17) and updates of $(\mathbf{h}_0,\mathbf{h}_1,\boldsymbol{\Phi})$ to solve (16).

To actually implement the procedure described above, we need estimates of $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\gamma$, and we must restrict the search space for the signals $(\mathbf{c}_0,\mathbf{c}_1)$ to a reasonable parametric set. In fact, since the desired signal structure will be computed adaptively at the receiver, the parameters describing the new signals must be communicated to the transmitter whenever the signal structure is updated. As a result, it is desirable to keep the dimensionality of the parameter space as small as possible. One possibility is to assign two linearly independent "basis" signature sequences $(\mathbf{s}_0,\mathbf{s}_1)$ to each user and adapt the signal structure by searching over the two-dimensional signal subspace spanned by the known basis signatures. This approach has an additional advantage related to estimating $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\gamma$. That is, an initial training period can be adopted prior to signal adaptation, during which we set $\mathbf{c}_0=\mathbf{s}_0$, $\mathbf{c}_1=-\mathbf{s}_0$ for the first half of the training period and $\mathbf{c}_0=\mathbf{s}_1$, $\mathbf{c}_1=-\mathbf{s}_1$ for the second half of the training period. $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\gamma$ can then be estimated from the received data (*i.e.*, blindly) during the training period in a straightforward manner. Once the training period is ended, signal adaptation can begin, and the estimates of $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\Sigma}_\gamma$ can be updated using a decision directed approach.

IV. SIMULATION RESULTS

We have completed a preliminary performance evaluation of the proposed adaptive signal/receiver scheme in the binary case. The results of that analysis are presented in this section. Throughout this section, we make the following assumptions:

- Both signals have equal energy, and the spreading gain is $N=31$ ($\|\mathbf{c}_0\|^2=\|\mathbf{c}_1\|^2=31$).
- The additive noise on the channel consists of ISI and AWGN only.
- The ISI is generated by frequency-selective Rayleigh fading with $L+1=16$ resolvable paths which are independent but not identically distributed. The fading parameters are assumed constant over blocks of 100 bits and independent from block to block. The total power of the fading process (*i.e.*, the total variance for all paths) is 10.

- The statistics of the channel are known *a priori*; in particular, the covariance matrix is diagonal and:

$$\begin{aligned}\mathbf{\Sigma}_{\alpha} &= \mathbf{\Sigma}_{\varepsilon} + \mathbf{\Sigma}_{\hat{\alpha}},\\ \mathbf{\Sigma}_{\varepsilon} &= \beta\mathbf{\Sigma}_{\alpha},\\ \mathbf{\Sigma}_{\hat{\alpha}} &= (1-\beta)\mathbf{\Sigma}_{\alpha},\\ \mathbf{\Sigma}_{\gamma} &= \sigma^2\mathbf{I},\end{aligned} \tag{18}$$

  where $0 \le \beta \le 1$ represents the *uncertainty coefficient* of the channel and $\sigma^2 > 0$ represents the variance of the additive channel noise (excluding ISI). A completely coherent channel corresponds to $\beta = 0$, and a completely noncoherent channel corresponds to $\beta = 1$. The channel SNR is given by $E_b/N_0 = 31/\sigma^2$.

The results of a sequence of simulations are presented in Figure 1 below. In this figure, the performance of the adaptive LQ receiver is compared with the performance of a linear MMSE receiver employing antipodal modulation. The signal sets for both the linear detector and the LQ detector were chosen adaptively to give the best performance in each case relative to the instantaneous CSI and the known channel statistics. For the linear detector, only $\mathbf{c}_0$ was chosen adaptively, since antipodal modulation ($\mathbf{c}_1 = -\mathbf{c}_0$) was assumed. For the LQ detector, both $\mathbf{c}_0$ and $\mathbf{c}_1$ were chosen adaptively subject to the equal-energy constraint.

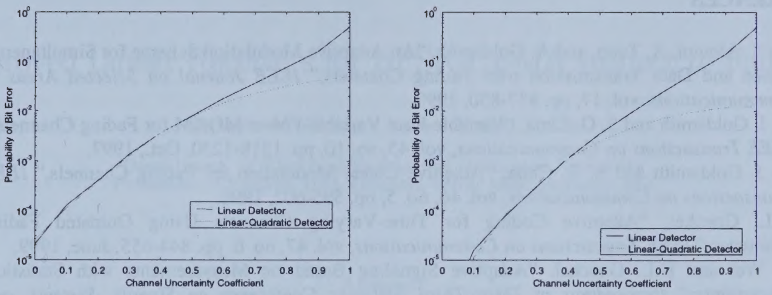
As the figure indicates, the performance of the adaptive LQ detector is always at least as good as the performance of the linear detector. This is the desired result and indicates not only that the modified deflection ratio is a good criterion for signal selection for this problem, but also that the recursive LMS algorithm discussed above converges to a nearly optimal solution. Furthermore, as one would expect, the LQ detector becomes strictly better than the linear detector as the uncertainty on the channel increases at a fixed SNR. Finally, as the SNR on the channel increases, the level of uncertainty at which the LQ detector begins to significantly outperform the linear detector converges to zero. This is consistent with the fact that as the SNR on the channel increases, the dominant error mechanism on the channel becomes the mismatch between the detector and the channel that results from inaccurate CSI estimates.

The results of the same sequence of simulations are summarized and plotted as a function of SNR in Figure 2. Here the increasing advantage of the LQ receiver relative to a desired bit error rate is clearly displayed. In particular, for values of $\beta > 0.2$, the performance gain associated with the LQ detector in the operating range of interest (say $P_e < 10^{-3}$) is dramatic.

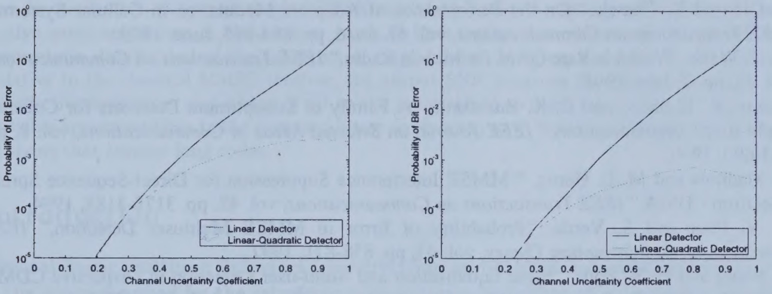
V. CONCLUSION

The results of this study indicate that LQ receivers can provide significant performance improvements in the presence of channel uncertainty on single-user channels at a modest cost in receiver complexity. The advantages for multiuser channels have not been demonstrated, but similar results can be expected. Many issues regarding the performance of LQ receivers remain to be studied. The following topics are currently under investigation:

- Asymptotic multiuser efficiency and near-far resistance of LQ receivers.
- Sensitivity of LQ receivers to errors in the estimates of relevant channel statistics.
- Analytical evaluation of receiver performance.
- Performance of adaptive implementations.



(a) 5 dB (b) 8 dB



(c) 11 dB (d) 14 dB

Figure 1. Performance of Linear vs. LQ Receiver – P_e vs. Channel Uncertainty

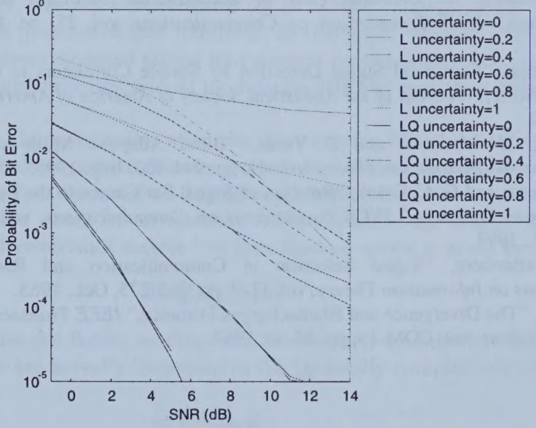


Figure 2. Performance of Linear vs. LQ Receiver – P_e vs. SNR

REFERENCES

[1] M. S. Alouini, X. Tang, and A. Goldsmith, "An Adaptive Modulation Scheme for Simultaneous Voice and Data Transmission over Fading Channels," *IEEE Journal on Selected Areas in Communications,* vol. 17, pp. 837-850, 1999.

[2] A. J. Goldsmith and S. G. Chua, "Variable-Rate Variable-Power MQAM for Fading Channels," *IEEE Transactions on Communications*, vol. 45, no. 10, pp. 1218-1230, Oct., 1997.

[3] A. J. Goldsmith and S. G. Chua, "Adaptive Coded Modulation for Fading Channels," *IEEE Transactions on Communications*, vol. 46, no. 5, pp. 595-602, 1998.

[4] D.L. Goeckel, "Adaptive Coding for Time-Varying Channels Using Outdated Fading Estimates," *IEEE Transactions on Communications*, vol. 47, no. 6, pp. 844-855, June, 1999.

[5] S. Wei and D.L. Goeckel, "Adaptive Signaling Based on Measurements with Statistical Uncertainty," *Proceedings of Thirty-Third Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, vol. 1, pp. 27-31, Oct. 24-27, 1999.

[6] X. Qiu and K. Chawla, "On the Performance of Adaptive Modulation in Cellular Systems," *IEEE Transactions on Communications*, vol. 47, no. 6, pp. 884-895, June, 1999.

[7] W. T. Webb, "Variable Rate QAM for Mobile Radio," *IEEE Transactions on Communications*, vol. 43, no. 7, July, 1995.

[8] Z. Xie, R. T. Short, and C. K. Rushforth, "A Family of Suboptimum Detectors for Coherent Multi-user Communications," *IEEE Journal on Selected Areas in Communications*, vol. 8, pp. 683-690, 1990.

[9] U. Madhow and M. L. Honig, " MMSE Interference Suppression for Direct-Sequence Spread Spectrum CDMA," *IEEE Transactions on Communications*, vol. 42, pp. 3178-3188, 1994.

[10] H. V. Poor and S. Verdu, "Probability of Error in MMSE Multiuser Detection," *IEEE Transactions on Information Theory*, vol. 43, pp. 858-871, 1997.

[11] X. Wang and H. V. Poor, "Blind Equalization and Multi-user Detection in Dispersive CDMA Channels," *IEEE Transactions on Communications*, vol. 46, pp. 91-103, Jan., 1998.

[12] X. Wang and H. V. Poor, "Blind Multiuser Detection: A Subspace Approach," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 677-690, March, 1998.

[13] R. J. Barton and H. V. Poor, "On Generalized Signal-to-Noise Ratios in Quadratic Detection," *Mathematics of Control, Signals, and Systems*, vol. 5, pp. 81-91, 1992.

[14] W. A. Gardner, "A Unifying View of Second-Order Measures of Quality for Signal Classification," *IEEE Transactions on Communications*, vol. 28, no. 6, pp. 807-816, June, 1980.

[15] D. Middleton, "Acoustical Signal Detection by Simple Correlators in the Presence of Non-Gaussian Noise," *Journal of the Acoustical Society of America of America*, vol. 34, pp. 1598-1609, 1962.

[16] M. Honig, U. Madhow, and S. Verdu, "Blind Adaptive Multi-user Detection," *IEEE Transactions on Information Theory*, vol. 41, pp. 944-960, July, 1995.

[17] D. H. Johnson and G. C. Orsak, "Relation of Signal Set Choice to the Performance of Optimal Non-Gaussian Detectors," *IEEE Transactions on Communications*, vol. 41, no. 9, pp. 1319-1328, Sept., 1993.

[18] T. L. Grettenberg, "Signal Selection in Communication and Radar Systems," *IEEE Transactions on Information Theory*, vol. IT-9, pp. 265-275, Oct., 1963.

[19] T. Kailath, "The Divergence and Bhattacharyya Distance," *IEEE Transactions on Communications*, vol. COM-15, pp. 52-60, 1967.

Improved Linear Receivers for BPSK-CDMA subject to Fading*

Antonia M. Tulino and Sergio Verdú

Department. of Electrical Engineering, Princeton University
Princeton, NJ 08544, USA {atulino, verdu }@ee.princeton.edu

Abstract

In this paper we design and analyze a new class of linear multiuser detectors, which can be used when the users employ BPSK modulation and the fading coefficients of the active users are known at the receiver (such as base-station demodulation). The tools of asymptotic distribution of the spectrum of large random matrices are used to show that relative to the classical MMSE receiver, the output SNR improves by halving the number of effective interferers and adding 3 dB to the input signal-to-noise ratio. We also propose sensible approximations to the proposed linear receivers so as to facilitate their use in CDMA systems that employ long codes.

1 Introduction

The classical linear multiuser detectors (decorrelating and MMSE receivers) exploit the geometry of the space spanned by the interference to mitigate the effect of multiaccess interference. However, unlike nonlinear multiuser detectors (maximum-likelihood, successive cancellation, etc.) those detectors do not exploit the structure of the modulation format of the desired user and the multiple-access interference. There is much to be gained by exploiting the structure of the modulation. For example, when an interferer is very strong a nonlinear receiver can essentially nullify its effect, whereas a decorrelator (or an MMSE receiver) would pay a penalty in performance due to the presence of that interferer. In this paper, we take a hybrid approach that considers linear receivers designed taking into account certain information available when the interferers use antipodal one-dimensional modulation (BPSK) and their channel fades are known at the receiver.

Unlike additive white Gaussian noise, when the users employ BPSK modulation, the baseband equivalent of the multiaccess interference process has a nonzero pseudo-autocovariance function when conditioned on the received fading coefficients. To see this, consider the contribution to the observable vector of an interferer with signature $\mathbf{s}_k$, symbol b_k and fading coefficient α_k. The pseudocovariance matrix [1] of that random vector is given by:

$$E[\alpha_k^2]E[b_k^2]\mathbf{s}_k\mathbf{s}_k^T$$

which is equal to 0 because the fading coefficient α_k is circularly symmetric. However, if α_k is known at the receiver, we are actually interested in the (generally complex valued) conditional pseudocovariance matrix

$$\alpha_k^2 E[b_k^2]\mathbf{s}_k\mathbf{s}_k^T.$$

If b_k were the result of a QPSK (or other such symmetric constellations in the complex plane), then $E[b_k^2] = 0$. However, in the case of BPSK $E[b_k^2] = 1$, and the pseudocovariance matrix of the multiaccess interferers whose fading values are known at the receiver is nonzero.

---

*This research is supported in part by CNR Grant No.23.07.00 and NFS Grant CCR-0074277

[1] Unlike the covariance matrix which averages a vector times its Hermitian, the pseudo covariance matrix is the average of the vector times its transpose

Following up on work of the first author in [2, 1], in this paper we exploit the nonzero conditional pseudocovariance functions to obtain alternative decorrelator and MMSE receivers. Since these linear transformations operate on both the received observable vector and its conjugate we refer to them as *linear conjugate receivers.*

Our new performance results of the linear conjugate receivers use the tools developed in recent years [6, 5, 7] for the analysis of multiuser detectors in the wideband limit when the number of users goes to infinity while keeping the ratio β of users to spreading-gain constant. The performance gain afforded over the classical linear multiuser receivers by the linear conjugate receivers can be quite dramatic. For example, the near-far resistance of the decorrelator becomes $1-\beta/2$ instead of $1-\beta$ and the output SNR of the MMSE depends on β and the input SNR in the same way as in the classical MMSE receiver (Tse-Hanly formula [5]) except that the effective number of interferers is halved and 3 dB is added to the input SNR.

Another contribution of this paper is to propose low-complexity approximations to the original linear-conjugate receivers, which would be of interest when the CDMA system employs long codes. Matrix inverses are approximated by polynomials whose coefficients can be computed in various ways. A particularly promising way is to obtain them through an asymptotic design which, like the asymptotic SNR analysis, uses recent results on the asymptotic eigenvalue distribution of random matrices.

2 System Model

We consider a synchronous BPSK-modulated DS/CDMA system with K active users, employing long spreading codes and operating over a frequency-flat fading channel. The baseband equivalent of the received signal is:

$$r(t) = \sum_{\ell=-\infty}^{\infty}\sum_{k=1}^{K}\alpha_k(\ell)\sqrt{\mathcal{E}_k}\,b_k(\ell)\,s_k^{\ell}(t-\ell T_b)+n(t) \qquad (1)$$

where

- $\mathcal{E}_k$ is the energy of the $k-$th user;
- $\{b_k(\ell)\}_{\ell=-\infty}^{+\infty}$ represents the bit stream of the $k-$th user, modeled as a sequence of independent and identically distributed binary variables, each taking on values in the set $\{-1,1\}$;
- T_b is the bit interval duration;
- $s_k^{\ell}(t)$ is the *signature* waveform assigned to the $k-$th user in the $\ell-$th signaling interval, expressed as

$$s_k^{\ell}(t)=\sum_{n=0}^{N-1} s_{k,n}^{\ell}\,u_{T_c}(t-nT_c)$$

with N the processing gain, $\boldsymbol{s}_k^{\ell}=[s_{k,0}^{\ell},\ldots,s_{k,N-1}^{\ell}]$ the $k-$th spreading code in the $\ell-$th signaling interval, T_c the chip interval, and $u_{T_c}(\cdot)$ is a chip waveform with zero autocorrelation at multiples of T_c and unit energy. The spreading code $\boldsymbol{s}_k^{\ell}$ is normalized to have unit norm and its elements[2] are $s_{k,n}^{\ell}\in\{-1/\sqrt{N},1/\sqrt{N}\}\quad n=0,\ldots,N-1$.

- $\alpha_k(\ell)$ is a complex channel gain. The complex-valued fading parameters $\alpha_k(\ell)$ are zero-mean independent from user to user and follow a common distribution (for all k and ℓ) such that $\mathrm{E}\left[\alpha_k(\ell)\alpha_n^*(\ell)\right]=\delta_{nk}$ and $\mathrm{E}\left[\alpha_k(\ell)\alpha_n(\ell)\right]=0$ with $(\cdot)^*$ denoting conjugate.

---

[2] Similar results can be obtained for non-binary sequences.

Let us assume that we want to demodulate user "1" , and that a decision about $b_1(\ell)$ is made by processing the observation in the interval $[\ell T_b, (\ell+1)T_b]$. Chip-matched filtering the received waveform in the interval $[\ell T_b, (\ell+1)T_b]$ we obtain the following N-dimensional vector sequence of observables:

$$\begin{aligned} \boldsymbol{r}(\ell) &= b_1(\ell)\sqrt{\mathcal{E}_1}\boldsymbol{s}_1^\ell \alpha_1(\ell) + \sum_{k=2}^{K} b_k(\ell)\sqrt{\mathcal{E}_k}\boldsymbol{s}_k^\ell \alpha_k(\ell) + \boldsymbol{n}(\ell) = \\ &= b_1(\ell)\sqrt{\mathcal{E}_1}\boldsymbol{s}_1^\ell \alpha_1(\ell) + \boldsymbol{S}_1^\ell \boldsymbol{A}_1 \boldsymbol{B}_1(\ell)\boldsymbol{\alpha}_1(\ell) + \boldsymbol{n}(\ell) \end{aligned} \tag{2}$$

where $\boldsymbol{A}_1 = \operatorname{diag}\left(\sqrt{\mathcal{E}_2}, \dots, \sqrt{\mathcal{E}_K}\right)$, $\boldsymbol{B}_1(\ell) = \operatorname{diag}\left(b_2(\ell), \dots, b_K(\ell)\right)$, $\boldsymbol{S}_1^\ell$ is a $N \times (K-1)$ dimensional matrix whose columns are the signatures $\left\{\boldsymbol{s}_k^\ell\right\}_{k=2}^{K}$ and $\boldsymbol{\alpha}_1(\ell) = [\alpha_2(\ell), \dots, \alpha_K(\ell)]$ is the $K-1$ dimensional fading coefficient vector. In (2) the first term on the right-hand-side (RHS) represents the contribution from the bit to be decoded, while the other terms represent the contributions from MAI ($\boldsymbol{z}(\ell)$) and the thermal noise ($\boldsymbol{n}(\ell)$), respectively. Finally, notice that $\boldsymbol{n}(\ell)$ is a white complex Gaussian vector with covariance matrix $2\mathcal{N}_0\boldsymbol{I}_N$, with $\boldsymbol{I}_N$ the identity matrix of order N. The covariance matrix of the observables conditioned on all fading variables has the following expression:

$$\boldsymbol{M}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell) \triangleq \mathrm{E}\left[\boldsymbol{r}(\ell)\boldsymbol{r}^H(\ell)|\boldsymbol{\alpha}_1(\ell)\right] = \mathcal{E}_1|\alpha_1(\ell)|^2\boldsymbol{s}_1^\ell\boldsymbol{s}_1^{\ell H} + \sum_{k=2}^{K}\mathcal{E}_k|\alpha_k(\ell)|^2\boldsymbol{s}_k^\ell\boldsymbol{s}_k^{\ell H} + 2\mathcal{N}_0\boldsymbol{I}_N \tag{3}$$

with $(\cdot)^H$ denoting conjugate transpose. The pseudocovariance matrix conditioned on all the fading variables [3] is given by:

$$\boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell) \triangleq \mathrm{E}\left[\boldsymbol{r}(\ell)\boldsymbol{r}^T(\ell)|\boldsymbol{\alpha}_1(\ell)\right] = \mathcal{E}_1(\alpha_1(\ell))^2\boldsymbol{s}_1^\ell\boldsymbol{s}_1^{\ell T} + \sum_{k=2}^{K}\mathcal{E}_k(\alpha_k(\ell))^2\boldsymbol{s}_k^\ell\boldsymbol{s}_k^{\ell T} \tag{4}$$

with $(\cdot)^T$ denoting transpose. Since $\boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}$ is nonzero, the baseband equivalent of the CDMA signals is an improper random processes[3]. This is also a straightforward, although nontrivial, consequence of the fact that, since the radio-frequency CDMA signals, conditioned on the fading coefficients, are not wide-sense stationary (WSS), the correlation properties of the corresponding complex envelopes are specified by four real functions, or, equivalently, by two complex functions. Likewise, the projections of these complex envelopes onto an orthonormal system are specified by two nonzero complex matrices, $\boldsymbol{M}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}$ and $\boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}$.

For future reference, we denote by $\mathcal{R}(\boldsymbol{A})$ the column space of $\boldsymbol{A}$, by $\mathcal{N}(\boldsymbol{A})$ the null space of $\boldsymbol{A}$, by $\mathcal{C}^N$ the space of $N-$tuples on the complex field $\mathcal{C}$ with the usual internal and external operations, and by $\mathcal{S}(\ell) \triangleq \mathcal{R}\left(\boldsymbol{S}_1^\ell\right)$ the *interference subspace* , namely the subspace of $\mathcal{C}^N$ spanned by the MAI (i.e. by $\{\boldsymbol{s}_k^\ell\}_{\ell,k}$ with $k = 2, \dots, K$).

3 Detector Design

Any *linear* one-shot detector for user "1" implements a decision rule based on the projection of $\boldsymbol{r}(\ell)$ along a given vector:

$$\widehat{b}_1(\ell) = \operatorname{sgn}\left[\Re\left\{\boldsymbol{c}_1^H(\ell)\boldsymbol{r}(\ell)\right\}\right] \tag{5}$$

where $\operatorname{sgn}(\cdot)$ denotes the signum function, $\Re\{\cdot\}$ denotes real part, while the vector $\boldsymbol{c}_1(\ell) \in \mathcal{C}^N$ is a suitable direction, dictated by some optimality criterion, by the complexity constraints and by the information available at the receiver.

---

[3] According to [3], a complex random process $n(t)$ is said to be proper if its pseudoautocorrelation function $R_n(t,u) = \mathrm{E}[n(t)n(u)]$ is zero $\forall t, u$, and it is said to be improper in the opposite case that $R_n(t,u)$ is non-zero.

The most popular optimization criteria lead to the decorrelator and the MMSE techniques. If the desired signal does not belong to $\mathcal{S}(\ell)$, the decorrelator can be obtained as the unique solution to the following constrained maximization problem [6]:

$$\widetilde{\boldsymbol{c}}_1(\ell) = \arg \max_{s.t. \boldsymbol{c}_1^H \boldsymbol{z}(\ell)=0} \left\{ \frac{\Re\left\{\alpha_1(\ell)\boldsymbol{c}_1^H \boldsymbol{s}_1^\ell\right\}}{\|\boldsymbol{c}_1\|} \right\} \tag{6}$$

Likewise, the classical MMSE detector may be obtained as the solution to the problem [6]:

$$\widetilde{\boldsymbol{c}}_1(\ell) = \arg\min \mathrm{E}\left[\left|b_1(\ell) - \boldsymbol{c}_1^H \boldsymbol{r}(\ell)\right|^2\right] \tag{7}$$

or, alternatively, to the following constrained minimization of the output interference plus thermal noise energy:

$$\widetilde{\boldsymbol{c}}_1(\ell) = \arg \min_{s.t. \boldsymbol{c}_1^H \boldsymbol{s}_1^\ell = 1} \mathrm{E}\left[\left|\boldsymbol{c}_1^H(\boldsymbol{z}(\ell) + \boldsymbol{n}(\ell))\right|^2\right] \tag{8}$$

In spite of their being usually treated separately, the decorrelator (6), and MMSE receiver (8), can be subsumed under a single framework. In fact, both of them are solutions to a single general non linear constrained-optimization problem, viz.,

$$\left\{ \begin{array}{l} \widetilde{\boldsymbol{c}}_1(\ell) = \arg \min\limits_{\boldsymbol{c}_1 \in \mathcal{D}(\ell)} \|\boldsymbol{c}_1\| \\[2ex] \mathcal{D}(\ell) \triangleq \left\{ \boldsymbol{c}_1 \in \mathcal{C}^N:\ \mathrm{E}_{\chi(\ell)}\left[\Re^2\left\{\boldsymbol{c}_1^H \boldsymbol{w}(\ell)\right\}\right] = \min,\ \ s.t.\ \Re\left\{\alpha_1(\ell)\boldsymbol{c}_1^H \boldsymbol{s}_1^\ell\right\} = 1 \right\} \end{array} \right. \tag{9}$$

for different choices of the vector $\boldsymbol{w}(\ell)$. In (9) $\chi(\ell)$ is the vector of all parameters in $\boldsymbol{w}(\ell)$ which are not known to the receiver, and with respect to which we take the expectation. Specifically, here we are interested in two choices:

1. $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell) \triangleq \boldsymbol{S}_1^\ell \boldsymbol{A}_1 \boldsymbol{B}_1(\ell) \boldsymbol{\alpha}_1(\ell)$
2. $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell) + \boldsymbol{n}(\ell)$

which yield, as the solution to (9), a family of decorrelators and a family of MMSE receivers, respectively. Since the expectation in (9) is with respect to all parameters in $\boldsymbol{w}(\ell)$ not known to the receiver, the solution to (9) depends on the prior information available at the receiver. To illustrate further, let us consider the following situations:

[**a**] The receiver has prior knowledge of the fading coefficients $\{\alpha_1(\ell)\}_\ell$, but not of the realizations of the fading coefficients of the other users which are modeled as random variables.

[**b**] The receiver has prior knowledge of the realizations of the fading coefficients of all users $\{\alpha_k(\ell)\}_{\ell,k}$.

In both situations we assume that in each signaling interval the receiver knows the signatures of all users. Finally notice that $\Re\left\{\boldsymbol{x}^H \boldsymbol{y}\right\}$ (with $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}^N$) is not linear in x and therefore we cannot apply the ortogonality principle. Thus, unlike the conventional optimization criteria (6) and (8), the constrained minimization problem (9) is not linearly constrained, and hence its solution requires special attention.

3.1 Unconditional linear receivers

First, let us focus on the situation [a]. In this case to solve the constrained minimization problem (9) is equivalent to solving:

$$\widetilde{\boldsymbol{c}}_1(\ell) = \arg \min_{s.t.\ \boldsymbol{c}_1 \in \mathcal{D}(\ell)} \|\boldsymbol{c}_1\| \tag{10}$$

with $\mathcal{D}(\ell) \equiv \left\{ \boldsymbol{c}_1 \in \mathcal{C}^N :\ \mathrm{E}_{\chi(\ell)} \left[\left| \boldsymbol{c}_1^H \boldsymbol{w}(\ell) \right|^2 \right] = \min,\ \ s.t.\ \Re\left\{ \alpha_1(\ell) \boldsymbol{c}_1^H \boldsymbol{s}_1^\ell \right\} = 1 \right\}$.

1. Assuming that $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell)$, the solution to (10) is given by the following expression:

   $$\boldsymbol{c}_{\mathrm{UD}}(\ell) = \gamma_{UD} \alpha_1(\ell) \left(\boldsymbol{M}_{\boldsymbol{zz}}(\ell) + \mathcal{E}_1 |\alpha_1(\ell)|^2 \boldsymbol{s}_1^\ell \boldsymbol{s}_1^{\ell H} \right)^+ \boldsymbol{s}_1^\ell \tag{11}$$

   where $\boldsymbol{M}_{\boldsymbol{zz}}(\ell) = E_{\chi(\ell)}[\boldsymbol{z}(\ell)\boldsymbol{z}(\ell)^H] = \boldsymbol{S}_1^\ell \boldsymbol{A}_1^2 \boldsymbol{S}_1^{\ell H}$ is the covariance matrix of the interference vector $\boldsymbol{z}(\ell)$ with $\chi(\ell) = [\boldsymbol{B}_1(\ell), \boldsymbol{\alpha}_1(\ell)]$, $(\cdot)^+$ denotes pseudo-inverse [6] and $\gamma_{\mathrm{UD}} = 1/(\alpha_1(\ell) \boldsymbol{s}_1^{\ell H} \boldsymbol{M}_{\boldsymbol{zz}}^+(\ell) \boldsymbol{s}_1^\ell)$. Notice that when $\boldsymbol{s}_1^\ell \notin \mathcal{S}(\ell)$, $\min \mathrm{E}\left[\boldsymbol{c}_1^H \boldsymbol{z}(\ell) \right]$ is equal to zero. As a consequence to solve the problem (10) for $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell)$ is equivalent to solving (6); thus (11) coincides with the classical Decorrelator detector given by (6). If instead, $\boldsymbol{s}_1^\ell \in \mathcal{S}(\ell)$, then (11) depends also on the energy of the interferers. We refer to (11) to as the *Unconditional* Decorrelator (UD).

2. When, instead, $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell) + \boldsymbol{n}(\ell)$, problem (10) admits as its solution the classical *Unconditional* MMSE (UM) receiver (8):

   $$\boldsymbol{c}_{\mathrm{UM}}(\ell) = \gamma_{\mathrm{UM}} \alpha_1(\ell) \boldsymbol{M}_{\boldsymbol{rr}}^{-1}(\ell) \boldsymbol{s}_1^\ell \tag{12}$$

   where $\boldsymbol{M}_{\boldsymbol{rr}}(\ell) = E_{\chi(\ell)}[\boldsymbol{r}(\ell)\boldsymbol{r}(\ell)^H]$ is the covariance matrix of the received vector $\boldsymbol{r}(\ell)$ with $\chi(\ell) = [\boldsymbol{B}_1(\ell), \boldsymbol{\alpha}_1(\ell), \boldsymbol{n}(\ell)]$ and $\gamma_{\mathrm{UM}} = 1/(\alpha_1(\ell) \boldsymbol{s}_1^{\ell H} \boldsymbol{M}_{\boldsymbol{rr}}^{-1} \boldsymbol{s}_1^\ell)$.

3.2 Conditional linear receivers

Let us now consider situation [b], wherein the receiver has prior knowledge of the *realizations* of all fading coefficients, rather than of their ensemble features only. This might be the case of a base station in a cellular network. Consider the following results [2, 1]:

Proposition 1 *Assuming that* $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell)$, *define the* $N \times 2N-$*dimensional matrices* $\boldsymbol{F} = [\boldsymbol{I}_N\, \boldsymbol{0}]$ *and* $\boldsymbol{F}' = [\boldsymbol{0}\, \boldsymbol{I}_N]$; *then the solution to the constrained problem (9) is written as:*

$$\boldsymbol{c}_{\mathrm{LCD}}(\ell) = \gamma_{\mathrm{LCD}} \boldsymbol{F} \left(\mathrm{Q}_{\boldsymbol{\alpha}_1}^{a}(\ell) \boldsymbol{F}^T \boldsymbol{s}_1^\ell \alpha_1(\ell) + \mathrm{Q}_{\boldsymbol{\alpha}_1}^{a*}(\ell) \boldsymbol{F}'^T \boldsymbol{s}_1^\ell \alpha_1^*(\ell) \right) \tag{13}$$

wherein $\mathrm{Q}_{\boldsymbol{\alpha}_1}^{a}(\ell)$ *is the projector onto* $\mathcal{N}\begin{pmatrix} \boldsymbol{M}_{\boldsymbol{zz}|\boldsymbol{\alpha}_1}(\ell) & \boldsymbol{M}'_{\boldsymbol{zz}|\boldsymbol{\alpha}_1}(\ell) \\ \boldsymbol{M}'^*_{\boldsymbol{zz}|\boldsymbol{\alpha}_1}(\ell) & \boldsymbol{M}^*_{\boldsymbol{zz}|\boldsymbol{\alpha}_1}(\ell) \end{pmatrix}$ *and* γ_{LCD} *ensures that* $\Re\left\{ \alpha_1(\ell) \boldsymbol{c}_{LCD}^H \boldsymbol{s}_1^\ell \right\} = 1$.

We refer to (13) as the Linear-Conjugate Decorrelator (LCD).

Proposition 2 *Assuming* $\boldsymbol{w}(\ell) = \boldsymbol{z}(\ell) + \boldsymbol{n}(\ell)$, *the solution to the constrained minimization problem (9) is given by:*

$$\boldsymbol{c}_{\mathrm{LCM}}(\ell) = \gamma_{\mathrm{LCM}} \alpha_1(\ell) \boldsymbol{H}_{\boldsymbol{r}}^{-1}(\ell) \boldsymbol{s}_1^\ell - \alpha_1^*(\ell) \boldsymbol{M}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}^{-1}(\ell) \boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell) (\boldsymbol{H}_{\boldsymbol{r}}^*(\ell))^{-1} \boldsymbol{s}_1^\ell \tag{14}$$

with:

$$\boldsymbol{M}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell) = \mathrm{E}_{\chi(\ell)}[\boldsymbol{r}(\ell)\boldsymbol{r}^H(\ell)|\boldsymbol{\alpha}_1(\ell)], \qquad \boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell) = \mathrm{E}_{\chi(\ell)}[\boldsymbol{r}(\ell)\boldsymbol{r}^T(\ell)|\boldsymbol{\alpha}_1(\ell)],$$

$$\boldsymbol{H}_{\boldsymbol{r}}(\ell) = \left(\boldsymbol{M}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell) - \boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell)\left(\boldsymbol{M}^*_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell)\right)^{-1}\boldsymbol{M}'^*_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell)\right),$$

$\chi(\ell) = [\boldsymbol{B}_1(\ell), \boldsymbol{n}(\ell)]$ *and* γ_{LCM} *such that* $\Re\left\{\alpha_1(\ell)\boldsymbol{c}_{LCM}^H \boldsymbol{s}_1^\ell\right\} = 1.$

We refer to (14) as the Linear-Conjugate MMSE (LCM) receiver. Notice that, unlike the situation [*a*], the proposed LCM receiver (14) does not coincide with the classical Conditional MMSE receiver with the same prior knowledge. The expression of the latter, apart from an irrelevant positive factor, is given by:

$$\boldsymbol{c}_{\mathrm{CM}}(\ell) = \alpha_1(\ell)\boldsymbol{M}^{-1}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell)\boldsymbol{s}_1^\ell \tag{15}$$

and can be obtained by solving (8) where the expectation in (8) is with respect to $\chi(\ell) = [\boldsymbol{B}_1(\ell), \boldsymbol{n}(\ell)]$. We refer to (15) as the Conditional MMSE (CM) receiver.

The receivers defined by (13) and (14) outperform the Unconditional Decorrelator (11) and the Unconditional and Conditional MMSE receiver, (12) and (15) respectively; the reason is that (13) and (14) fully exploit the information about the fading coefficients resorting to Linear/Conjugate processing, wherein not only the received vector but also its conjugate are processed.

Before proceeding further in our discussion, we provide some intuition as to why Linear-Conjugate techniques cannot be worse than classical linear optimization techniques, (6) and (8). The key point is that both decorrelator and MMSE receivers are *linear* receivers, achieving interference elimination by projecting the observables onto suitable subspaces. On the other hand, if the fading vector is known, the interference vector $\boldsymbol{z}(\ell)$ and the received vector $\boldsymbol{r}(\ell)$ do *not* represent complex envelopes of stationary radio-frequency signals. As a consequence, their correlation properties are described by a pair of complex covariance matrices: such information is *not* contained in the conditional covariance matrix of the vectors themselves, but in the pair of conditional matrices: $\boldsymbol{M}_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell)$ and $\boldsymbol{M}'_{\boldsymbol{rr}|\boldsymbol{\alpha}_1}(\ell)$. Notice also that the unconditional pseudocovariance matrices $\mathrm{E}\left[\boldsymbol{z}(\ell)\boldsymbol{z}^T(\ell)\right]$ and $\mathrm{E}\left[\boldsymbol{r}(\ell)\boldsymbol{r}^T(\ell)\right]$ are zero (due to $\mathrm{E}\left[\alpha_k(\ell)\alpha_n(\ell)\right] = 0$), in keeping with the fact that, in the *ensemble* of the fading realizations, the vectors $\boldsymbol{z}(\ell)$ and $\boldsymbol{r}(\ell)$ *do* represent complex envelopes of stationary signals, whose correlation properties can thus be characterized through the conventional covariance matrix only. Thus, the conditional covariance matrix of the augmented observables contains additional information which is lost once an ensemble average is performed. It is worth noticing that the baseband solutions (13) and (14), (unlike the conventional (11) and (15)) achieve the same performance that radio-frequency (or intermediate-frequency) processing would be able to achieve.

3.2.1 Suboptimal receivers

Notice that both proposed receivers, (13) and (14), depend on the eigenvalues and eigenvectors of the covariance and pseudocovariance matrices of the interference vector $\boldsymbol{z}(\ell)$ conditioned on $\{\alpha_k(\ell)\}_{\ell,k}$. However, in long-code CDMA, the spreading codes as well as the corresponding matrices $\boldsymbol{M}_{\boldsymbol{zz}}(\ell)$, $\boldsymbol{M}_{\boldsymbol{zz}|\boldsymbol{\alpha}_1}(\ell)$ and $\boldsymbol{M}'_{\boldsymbol{zz}|\boldsymbol{\alpha}_1}(\ell)$, change in every symbol interval. This implies that the proposed optimal solutions, (13) and (14), have to be computed and updated symbol by symbol, thus leading to a prohibitive computational effort in most cases. As a consequence, a less complex approach is desirable. To understand how to simplify the optimum structure, let us define the subset of $\mathcal{C}^{2N}$:

$$\mathcal{S} \triangleq \left\{\boldsymbol{x}_a : \boldsymbol{x}_a = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{x}^* \end{pmatrix}, \boldsymbol{x} \in \mathcal{C}^N\right\}$$

It is easily seen that, with the following external operation (multiplication by the complex scalar γ):

$$\gamma \cdot \boldsymbol{x}_a = \begin{pmatrix} \gamma \boldsymbol{x} \\ \gamma^* \boldsymbol{x}^* \end{pmatrix}$$

and the usual component-wise vector sum as internal operation, $\mathcal{V} = \{\mathcal{S}, +, \cdot\}$ is a vector space on $\mathcal{C}$. Furthermore:

$$d^2(\boldsymbol{x}_a, \boldsymbol{y}_a) \triangleq \| \boldsymbol{x}_a - \boldsymbol{y}_a \|^2 = 2d^2(\boldsymbol{x}, \boldsymbol{y}) \qquad 2\Re\left\{\boldsymbol{x}^H \boldsymbol{y}\right\} = \boldsymbol{x}^H \boldsymbol{y} + \boldsymbol{x}^T \boldsymbol{y}^* = \boldsymbol{x}_a^H \boldsymbol{y}_a \tag{16}$$

then, by considering the new vector space $\mathcal{V}$ in lieu of $\mathcal{C}^N$ in (9), we can work with a linear constraint. As a consequence, define:

$$\mathbf{u}_{1a}^{\ell} = \begin{bmatrix} \alpha_1 \boldsymbol{s}_1^{\ell} \\ \alpha_1^* \boldsymbol{s}_1^{\ell} \end{bmatrix}, \quad \boldsymbol{U}_{1a}^{\ell} = \begin{bmatrix} \boldsymbol{S}_1^{\ell} \operatorname{diag}(\boldsymbol{\alpha}_1(\ell)) \\ \boldsymbol{S}_1^{\ell} \operatorname{diag}(\boldsymbol{\alpha}_1^*(\ell)) \end{bmatrix}, \tag{17}$$

$$\boldsymbol{M}_{\boldsymbol{z}_a \boldsymbol{z}_a | \boldsymbol{\alpha}_1}(\ell) = \mathrm{E}_{\chi(\ell)}\left[\boldsymbol{z}_a(\ell)\boldsymbol{z}_a^H(\ell)|\boldsymbol{\alpha}_1(\ell)\right] = \begin{pmatrix} \boldsymbol{M}_{\boldsymbol{z}\boldsymbol{z}|\boldsymbol{\alpha}_1}(\ell) & \boldsymbol{M}'_{\boldsymbol{z}\boldsymbol{z}|\boldsymbol{\alpha}_1}(\ell) \\ \boldsymbol{M}'^*_{\boldsymbol{z}\boldsymbol{z}|\boldsymbol{\alpha}_1}(\ell) & \boldsymbol{M}^*_{\boldsymbol{z}\boldsymbol{z}|\boldsymbol{\alpha}_1}(\ell) \end{pmatrix} \tag{18}$$

with $\chi = [\boldsymbol{B}_1(\ell), \boldsymbol{n}(\ell)]$ and denote by $\boldsymbol{U}_J^{\ell}$ the $2N \times J$ matrix whose columns form a linearly independent subset spanning the column space of matrix $\boldsymbol{U}_{1a}^{\ell}$. Then, with the definition $P_1(\ell) \triangleq \mathcal{E}_1|\alpha_1(\ell)|^2$, the solutions (13) and (14) to the constrained optimization problem (9) can be rewritten, except for an irrelevant positive factor, as:

$$\boldsymbol{c}_{\mathrm{LCD}} = \boldsymbol{F}\Big(P_1(\ell)\mathbf{u}_{1a}^{\ell}\mathbf{u}_{1a}^{\ell H} + \boldsymbol{M}_{\boldsymbol{z}_a\boldsymbol{z}_a|\boldsymbol{\alpha}_1}(\ell)\Big)^{\dagger} \mathbf{u}_{1a}^{\ell} = \boldsymbol{F}\left(\boldsymbol{I}_{2N} - \boldsymbol{U}_J^{\ell}(\boldsymbol{U}_J^{\ell H}\boldsymbol{U}_J^{\ell})^{-1}\boldsymbol{U}_J^{\ell H}\right) \boldsymbol{s}_{1a}^{\ell} \tag{19}$$

$$\boldsymbol{c}_{\mathrm{LCM}} = \boldsymbol{F}\Big(\boldsymbol{M}_{\boldsymbol{z}_a\boldsymbol{z}_a|\boldsymbol{\alpha}_1}(\ell) + 2\mathcal{N}_0 \boldsymbol{I}_{2N}\Big)^{-1} \mathbf{u}_{1a}^{\ell} \tag{20}$$

where $\boldsymbol{I}_{2N}$ is the $2N$-dimensional identity matrix. Based on (19) and (20), the $\boldsymbol{c}_{LCD}$ and the $\boldsymbol{c}_{LCM}$ vectors (13) and (14) can be expanded in an infinite power series of the conditional covariance matrix of the $2N-$dimensional vector $\boldsymbol{z}_a(\ell)$. As a consequence, the suboptimum approximation that we advocate consists of choosing $\widetilde{\boldsymbol{c}}_1(\ell)$ as corresponding to the minimum of (9) over a feasible constrained set. In this case after choosing the integer D depending on the allowable complexity, we obtain the parametric solutions:

$$\boldsymbol{c}_{\mathrm{SCD}}(\ell) = \sum_{m=0}^{D-1} w_{\mathrm{SCD}}^{m}(\ell)\boldsymbol{U}_J^{\ell}\left(\boldsymbol{U}_J^{\ell H}\boldsymbol{U}_J^{\ell}\right)^m \boldsymbol{U}_J^{\ell H}\mathbf{u}_{1a}^{\ell} \tag{21}$$

$$\boldsymbol{c}_{\mathrm{SCM}}(\ell) = \sum_{m=0}^{D-1} w_{\mathrm{SCM}}^{m}(\ell)\left(\boldsymbol{M}_{\boldsymbol{z}_a\boldsymbol{z}_a|\boldsymbol{\alpha}_1}(\ell) + 2\mathcal{N}_0\boldsymbol{I}_{2N}\right)^m \mathbf{u}_{1a}^{\ell} \tag{22}$$

where the vector weights, $\boldsymbol{w}_{(\cdot)}(\ell) = [w_{(\cdot)}^0(\ell), \ldots, w_{(\cdot)}^{D-1}(\ell)]$, are obtained by solving the constrained minimization problem (9). The optimal weight vectors are given by:

$$\boldsymbol{w}_{\mathrm{SCM}}(\ell) = \begin{pmatrix} \mathcal{H}_1^{\ell}, & \ldots, & \mathcal{H}_D^{\ell} \\ \vdots, & \ldots, & \vdots \\ \mathcal{H}_D^{\ell}, & \ldots, & \mathcal{H}_{(2D-1)}^{\ell} \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{H}_1^{\ell} \\ \vdots \\ \mathcal{H}_D^{\ell} \end{pmatrix}, \quad \boldsymbol{w}_{\mathrm{SCD}}(\ell) = \begin{pmatrix} \mathcal{Q}_1^{\ell}, & \ldots, & \mathcal{Q}_D^{\ell} \\ \vdots, & \ldots, & \vdots \\ \mathcal{Q}_D^{\ell}, & \ldots, & \mathcal{Q}_{(2D-1)}^{\ell} \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{Q}_1^{\ell} \\ \vdots \\ \mathcal{Q}_D^{\ell} \end{pmatrix} \tag{23}$$

where

$$\mathcal{H}_m^{\ell} \triangleq \mathbf{u}_{1a}^{\ell H}\left(\boldsymbol{M}_{\boldsymbol{z}_a\boldsymbol{z}_a|\boldsymbol{\alpha}_1}(\ell) + 2\mathcal{N}_0\boldsymbol{I}_{2N}\right)^m \mathbf{u}_{1a}^{\ell}, \qquad \mathcal{Q}_m^{\ell} \triangleq \mathbf{u}_{1a}^{\ell H}\left(\boldsymbol{U}_J^{\ell}\boldsymbol{U}_J^{\ell H}\right)^{m+2} \mathbf{u}_{1a}^{\ell} \tag{24}$$

We refer to (21) and (22) as Suboptimal Linear-Conjugate Decorrelator (SCD) and Suboptimal Linear-Conjugate MMSE detector (SCM) respectively. Eq. (23) implies that in order to compute the optimal weights (i.e. those solving the constrained minimization problem (9)) we have to invert a $D \times D$ Hankel matrix. The computational effort for on line inversion of the matrix in (23) is feasible when D is very small (this is equivalent, in order to obtain good performance, to saying that the number of users is not too large). If D is not too small the weight optimization approach cannot be adopted but we can resort to the design guidelines proposed at the end of this paper.

4 Performance Analysis

4.1 Nonasymptotic Analysis

Because of space limitation we omit our analytical results on nonasymptotic BER, Near-Far Resistance and signal-to-noise ratio performance. Our numerical results focus on a CDMA system employing Gold codes with spreading length $N = 31$ operating in Rayleigh fading. In Figure 1 we have represented the error probability versus the average received signal-to-noise ratio $\mathsf{SNR}= (\mathcal{E}_1)/(2\mathcal{N}_0)$, for the conditional MMSE receiver (15), labeled as "CM", for the newly proposed MMSE receiver (14), labeled as "LCM", and for the suboptimal linear-conjugate MMSE detector (22), labeled as "SCM" with $D = 10$. We assume 24 interfering users with a power level of 5dB above the desired signal. The results clearly show the superiority of the new strategy, which largely outperforms the conditional MMSE receiver (15). For comparison purposes, we also report the error probability corresponding to an uncoded BPSK transmission over a MAI-free flat-fading channel. Figure 2 contrasts the classical linear detection structures to the newly proposed receivers in terms of near-far resistance, which is represented versus K. Again, it is seen that the new approach allows a noticeable performance improvement.

4.2 Asymptotic Analysis

In the case of long sequences (and in other cases, e.g., with sequences distorted by random multipath, or in a network with random access) it is reasonable to assume that the spreading sequences are randomly and independently chosen. In this case, the performance measures of the receivers can be modeled as random variables, since they are a function of the spreading sequences. In the following we average the performance measures with respect to the random sequences and we also show that in the limiting regime $K \to \infty$, the random performance measures converge to deterministic quantities. The elements of the signatures $s_{k,n}^{\ell} \in \{-1/\sqrt{N}, 1/\sqrt{N}\}$ $\quad n = 0, \dots, N-1$ are chosen equally likely and independently for all (k, n, ℓ). Notice that similar results can be obtained for nonbinary sequences whose elements are independent and identically distributed (i.i.d.), with zero mean and variance $1/N$. The normalization ensures that $E[\|s_k^{\ell}\|] = 1$; however in this more general case we need also the assumption $E[|s_{k,n}^{\ell}|^4] < \infty$. The ratio of the number of users to the number of dimensions is denoted by $\beta \triangleq K/N$.

It is well known [6, 7] that the Near-Far resistance $\eta_1^C(\ell)$ of detectors (11), (12) and (15) converges almost surely as $K \to \infty$ to:

$$\lim_{K\to\infty} \eta_1^C(\ell) = [1-\beta]^+ \tag{25}$$

Based on some tools from free-probability theory [8] as applied to random matrices, we prove the following:

Proposition 3 *The average of $\eta_1(\ell)$ w.r.t. the signatures of all users is lower bounded by:*

$$E[\eta_1(\ell)] \geq \left[1 - \frac{K-1}{2N}\right]^+$$

Furthermore, the Near-Far resistance $\eta_1(\ell)$ for any $0 \leq \beta \leq \infty$ converges almost surely to:

$$\eta_1^{\infty} = \lim_{K \to \infty} \eta_1(\ell) = \left[1 - \frac{\beta}{2}\right]^{+} \tag{26}$$

Let us focus now on the (finite SNR) multiuser efficiency [6, 7] of the proposed receivers. Define a nonnegative random variable $|A|^2$ whose distribution is the limit distribution of $\left\{\frac{\mathcal{E}_k}{\mathcal{E}_1}|\alpha_k|^2,\ k = 1, \ldots K\right\}$. For the classical linear detection structures (11), (12) and (15), the multiuser efficiency for $K \to \infty$ converges almost surely to the solution $\eta_{CM}^{\infty}(\mathsf{SNR})$ of [5, 4]:

$$\eta + \beta \mathrm{E}\left[\frac{|A|^2\ \mathsf{SNR}\,\eta}{1 + |A|^2\ \mathsf{SNR}\,\eta}\right] = 1 \tag{27}$$

One of our main results is:

Proposition 4 *The multiuser efficiency of the Linear-Conjugate receivers (13) and (14) converges almost surely to the solution $\eta_{LCM}^{\infty}(\mathsf{SNR})$ of:*

$$\eta + \frac{\beta}{2} E\left[\frac{2\,\mathsf{SNR}\,|A|^2\ \eta}{1 + 2\,\mathsf{SNR}\,|A|^2\ \eta}\right] = 1 \tag{28}$$

The main conclusions from Propositions 3 and 4 are (i) the linear conjugate filters half the number of effective interferers and (ii) they add 3 dB to the input SNR. Effect (i) can be explained by noticing that the conventional decorrelator combats each interfering user as if it were quadrature-modulated which has the same noise-enhancement effect as having two BPSK-modulated users (for high SNR). To explain effect (ii) it is best to consider the single-user case; whereas the conventional MMSE filter minimizes the averaged magnitude squared at its output, the linear conjugate MMSE detector targets its resources to minimize the averaged real part squared (to exploit the BPSK modulation format) which has the same effect as turning off half of the input noise power.

In Figure 3 we show the asymptotic multiuser efficiency of the Linear-Conjugate MMSE receiver versus the ratio $\beta = K/N$ for a fixed average received signal-to noise ratio $\mathsf{SNR} = 10dB$ and for equal powers. For comparison purposes the spectral efficiency of the conditional MMSE receiver (CM) for equal powers [4] is also given.

So far we have only analyzed the unconstrained-complexity solutions to (9). Let us focus on the suboptimal structure receivers (21) and (22) that we obtain as solution to (9) when complexity constraints are imposed. As we have already noticed, the optimal weights (23) can be computationally intensive. Instead we propose an asymptotic weighting that assumes to operate in a limiting regime (i.e. the number of users and processing gain are large with fixed ratio and random spreading sequences). Using some recent results on the asymptotic eigenvalues distribution of random matrices we have:

$$\lim_{K \to \infty} \boldsymbol{s}_1^{\ell H}\left(\boldsymbol{M}_{\boldsymbol{z}_a \boldsymbol{z}_a|\boldsymbol{\alpha}_1}(\ell) + 2\mathcal{N}_0 \boldsymbol{I}_{2N}\right)^m \boldsymbol{s}_1^{\ell} \overset{a.s.}{=} E\left[\mathrm{trace}\left(\boldsymbol{M}_{\boldsymbol{z}_a \boldsymbol{z}_a|\boldsymbol{\alpha}_1} + 2\mathcal{N}_0 \boldsymbol{I}_{2N}\right)^m\right]$$

$$\lim_{K \to \infty} \boldsymbol{s}_1^{\ell H}\left(\boldsymbol{M}_{\boldsymbol{z}_a \boldsymbol{z}_a|\boldsymbol{\alpha}}(\ell)\right)^m \boldsymbol{s}_1^{\ell} \overset{a.s.}{=} E\left[\mathrm{trace}\left(\boldsymbol{M}_{\boldsymbol{z}_a \boldsymbol{z}_a|\boldsymbol{\alpha}_1}\right)^m\right]$$

It follows that:

$$\boldsymbol{w}_{\mathrm{SCD}}^{\infty} = \mathrm{E}\begin{bmatrix} \Lambda^3 & \Lambda^4, & \ldots, & \Lambda^{D+2} \\ \Lambda^4 & \Lambda^5, & \ldots, & \Lambda^{D+3} \\ \vdots & \vdots, & \ldots, & \vdots \\ \Lambda^{D+2} & \Lambda^{D+1}, & \ldots, & \Lambda^{2D+1} \end{bmatrix}^{-1} \mathrm{E}\begin{pmatrix} \Lambda^3 \\ \Lambda^4 \\ \vdots \\ \Lambda^{D+2} \end{pmatrix} \tag{29}$$

$$\boldsymbol{w}^{\infty}_{\text{SCM}} = \mathrm{E}\begin{bmatrix} (\Lambda+2\mathcal{N}_0) & (\Lambda+2\mathcal{N}_0)^2, & \ldots, & (\Lambda+2\mathcal{N}_0)^D \\ (\Lambda+2\mathcal{N}_0)^2 & (\Lambda+2\mathcal{N}_0)^3, & \ldots, & (\Lambda+2\mathcal{N}_0)^{D+1} \\ \vdots & \vdots, & \ldots, & \vdots \\ (\Lambda+2\mathcal{N}_0)^D & (\Lambda+2\mathcal{N}_0)^{D+1}, & \ldots, & (\Lambda+2\mathcal{N}_0)^{2D-1} \end{bmatrix}^{-1} \mathrm{E}\begin{pmatrix} (\Lambda+2\mathcal{N}_0) \\ (\Lambda+2\mathcal{N}_0)^2 \\ \vdots \\ (\Lambda+2\mathcal{N}_0)^D \end{pmatrix} \tag{30}$$

where the expectations are with respect to Λ, whose distribution is the asymptotic eigenvalue distribution of $\boldsymbol{M}_{\boldsymbol{z}_a\boldsymbol{z}_a|\boldsymbol{\alpha}_1}$. The entries of the matrices given in (29) and (30) can be computed by applying some tools of combinatorial theory [9]. In fact it can be shown that:

$$E\left[\Lambda^m\right] = \sum_{n=1}^{m} \beta^n \sum_{(j_1,\ldots,j_n)\in Z_n^{(m)}} \mathcal{E}_1^n c(j_1,\ldots,j_n) \mathrm{E}\left[(|A|^2)^{j_1}\right]\cdots \mathrm{E}\left[(|A|^2)^{j_n}\right]\cdot\beta^n$$

where the sum is over all possible $j_1,\ldots,j_n > 0$ such that $j_1+\cdots j_n = n$ and for all $1 \le n \le m$ [9]. We refer to (21) and (22) (where the weights are given by (29) and (30)) as the Asymptotic Suboptimal Linear-Conjugate Decorrelator (ASD) and Asymptotic Suboptimal Linear-Conjugate MMSE detector (ASM). In Figure (3) we have represented the error probability versus the average received signal-to-noise ratio SNR= $(\mathcal{E}_1)/(2\mathcal{N}_0)$, for the newly proposed MMSE receiver (14), labeled as "LCM", and for the asymptotic suboptimal linear-conjugate MMSE detector (22), labeled as "ASM" with $D = 15$. We assume $N = 150$ and $K = 30$ and interfering users with equal powers. For comparison purposes, we also report the error probability corresponding to an uncoded BPSK transmission over a MAI-free flat-fading channel.

References

[1] S. Buzzi, M. Lops, A. Tulino, "A New Class of Multiuser CDMA Receivers Based on The Minimum Mean-Output-Energy Strategy", *IEEE International Symposium on Information Theory*, Sorrento, Italy, 25-30 June 2000.

[2] M. Lops, A.M. Tulino,"Linear and Linear-Conjugate/Linear (LC/L) Receivers for CDMA Systems in Fading Channels", *IEEE 6th Int. Symp. on Spread Spectrum Tech. & Appl.*, New Jersey, USA, September 6-8, 2000.

[3] F.D. Neeser and J.L. Massey, "Proper Complex Random Processes with Applications to Information Theory", *IEEE Trans. on Information Theory*, Vol. 39, pp. 1293-1302, July 1993.

[4] S. Shamai and S. Verdú, "The Effect of Frequency-Flat Fading on the Spectral Efficiency of CDMA", to appear on IEEE Trans. on Information Theory.

[5] D.N.C.Tse and S.V. Hanly, " Linear multiuser receivers: effective interference, effective bandwidth and user capacity". *IEEE Trans. on Information Theory*, Vol. 45 2, pp.641-657, March 1999.

[6] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.

[7] S. Verdu and S. Shamai, "Spectral Efficiency of CDMA with Random Spreading" *IEEE Transactions on Information Theory*, Vol. 45 2, pp.622-640 March 1999.

[8] D. Voiculescu, K.J. Dykema and A. Nica, "Free Random Variables", ser. CRM Monograph. Providence, RI:Amer. Math. Soc., 1992, Vol.1.

[9] W. Xiao and M. L. Honig, "Convergence Analysis of Adaptive Full-Rank and Multi-Stage Reduced-Rank Interference Suppression", *2000 Conference on Information Sciences and Systems*, Princeton, March 15-17, 2000.

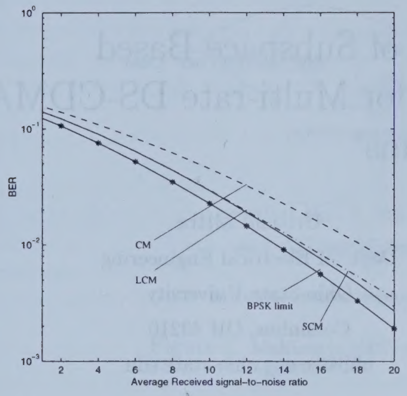


Figure 1: BER of the classical and of the newly proposed MMSE detectors, LCM and SCM, for $K = 25$.

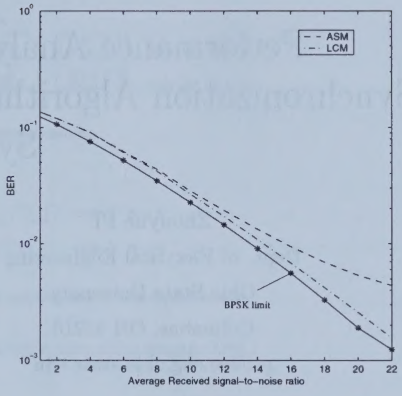


Figure 3: BER of the classical and of the newly proposed MMSE detectors, LCM and ASM, for $K = 30$ and $N = 150$.

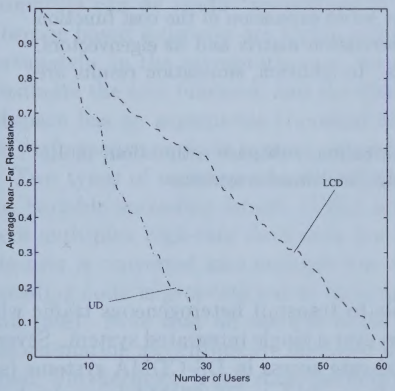


Figure 2: Near-Far resistance of the classical decorrelator and of the newly proposed Linear-Conjugate decorrelator.

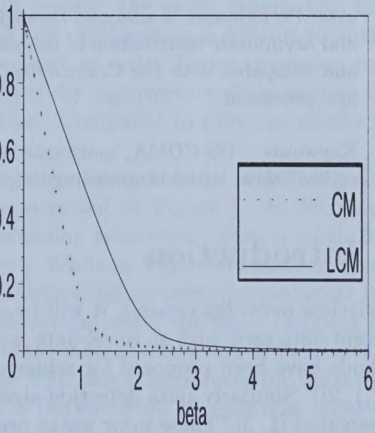


Figure 4: LCM multiuser efficiency for SNR = 10dB.

Performance Analysis of Subspace Based Synchronization Algorithms for Multi-rate DS-CDMA Systems*

Zhouyue Pi
Dept. of Electrical Engineering
Ohio State University
Columbus, OH 43210
piz@ee.eng.ohio-state.edu

Urbashi Mitra
Dept. of Electrical Engineering
Ohio State University
Columbus, OH 43210
ubli@ee.eng.ohio-state.edu

Abstract

The performance of subspace based delay estimation algorithms for multi-rate asynchronous DS-CDMA systems is investigated. Two multi-rate access schemes, multi-code (MC) access and variable-spreading-length (VSL) access, are considered. A subspace based algorithm that can explicitly exploit the redundant information introduced by the multi-rate nature of the signal is proposed. A two stage algorithm with reduced complexity and strong performance is proposed for the VSL case. Performance is analyzed through Taylor series expansion of the cost function and asymptotic distribution of the sample correlation matrix and its eigenvectors, and compared with the Cramér-Rao Bounds. In addition, simulation results are also presented.

Keywords : DS-CDMA, multi-rate, synchronization, subspace estimation, multi-code CDMA, variable spreading length CDMA, multimedia systems

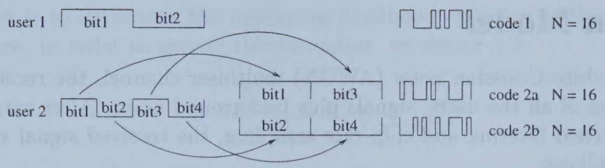
1 Introduction

As wireless networks expand, it will be desirable to transmit heterogeneous traffic with different data rate, such as voice, data, and video over a single integrated system. Several methods have been proposed for achieving multi-rate access in DS-CDMA systems (see *e.g.* [1, 2]). Similarly, data detection algorithms for multi-rate DS-CDMA have also been investigated [1, 3]. These prior works presumed perfect synchronization.

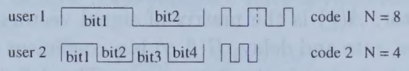
While training based synchronization algorithms have been considered in the past for single-rate systems (see *e.g.* [4, 5]), our focus will be on blind timing estimation algorithms. Blind algorithms can be more bandwidth efficient as data bits are not wasted on training. Blind methods based on second-order statistics have been widely considered for synchronization [6] and channel identification (see *e.g.* [7, 8]) in DS-CDMA systems. We shall extend such algorithms to the multi-rate scenario. It is noted that single-rate type algorithms [6, 9] were applied to multi-rate DS-CDMA signals in [10]; however,

---

[*] This work was supported by the National Science Foundation under CAREER Grant NCR 96-24375.



(a) Multi-code multi-rate communications



(b) Variable spreading length multi-rate communications with repitition codes

Figure 1: Multi-code (MC) and variable spreading length (VSL)

the multi-rate nature of the signal was not exploited. The focus of [10] was the use of improved correlation information.

In this paper, an improved projection of signal vectors onto the noise subspace based on maximum likelihood methods is presented. The subspace estimation based on this projection can explicitly exploit the multi-rate structure of the signal and provide better performance than previous MUSIC based estimators. We generalize the improved subspace based algorithm to a form from which previously proposed subspace-based timing estimators can be readily specialized [6, 10]. In general, the exact distribution of the subspace based estimates are highly complicated and performance is difficult to analyze. Fortunately, in the asymptotic case, we can use the first order Taylor expansion to approximate the cost function; and the distribution of the eigenvectors of the sample noise subspace has an asymptotic Gaussian distribution. Compared to previous results [11], the performance analysis is more accurate and general.

Two types of multi-rate access schemes will be considered: multi-code (MC) access and variable spreading length (VSL) access, as depicted in Figure 1. In MC access, users multiplex high-rate data onto low-rate spreading sequences. Thus a single high-rate user is converted into multiple low-rate users. While in VSL, the length of a user's spreading code is proportional to its symbol duration (and inversely proportional to its data rate). Note that all users in a MC or VSL system share the same chipping rate. The estimation algorithms are modified to exploit the inherent structure of the multi-rate signal. Furthermore, a two stage estimator for the VSL case is introduced for the purpose of complexity reduction, which takes advantage of the structure of repetition codes that we shall assume for the low rate VSL users.

This paper is organized as follows: in Section II, the system model for multi-rate CDMA systems is described. General results on subspace-based algorithms applicable to both MC and VSL multi-rate CDMA systems are presented in Section III. In addition, a two-stage estimation-detection scheme is proposed for delay estimation in VSL case in this section. Section IV provides the analytical derivation of performance of the proposed algorithms. In Section V, numerical examples are considered. Final conclusions are drawn in Section VI.

2 System Model

In an additive white Gaussian noise (AWGN) multiuser channel, the received signal is the superposition of all the users' signals plus background noise. After carrier demodulation, chip-matched filtering and chip-rate sampling, the received signal vector can be represented as follows:

$$\underline{r}(m) = A(\underline{\tau})B(\underline{\theta},\underline{\gamma})\underline{z}(m) + \underline{n}(m) \tag{1}$$

where $\underline{\tau}$, $\underline{\theta}$, $\underline{\gamma}$ are parameter vectors of all users' delay, phase, and amplitude, respectively; $\underline{r}(m)$ is the received vector; $A(\underline{\tau})$ is the matrix of signal vectors, which are function of the user's spreading codes, rate and delay. $B(\underline{\theta},\underline{\gamma})$ is a coefficient matrix, *i.e.*, the matrix for amplitude and phase; and $\underline{z}(m)$ is the data vector. The definition of these quantities depends on the particular multi-rate scheme employed. Due to the mutual independence of every user's signals, the matrices above have the following general form:

$$\begin{aligned} A(\underline{\tau}) &= \left[\begin{array}{cccc} A_1(\tau_1) & A_2(\tau_2) & \ldots & A_K(\tau_K) \end{array} \right] \\ B(\underline{\theta},\underline{\gamma}) &= diag\left[\begin{array}{ccc} B_1(\theta_1,\gamma_1) & \ldots & B_K(\theta_K,\gamma_K) \end{array} \right] \\ \underline{z}(m) &= \left[\begin{array}{cccc} \underline{z}_1^T(m) & \underline{z}_2^T(m) & \ldots & \underline{z}_K^T(m) \end{array} \right]^T \end{aligned} \tag{2}$$

where A_k is the matrix of signal vectors depending on τ_k (delay of the kth user); B_k is the diagonal coefficient matrix of the kth user; and $\underline{z}_k(m)$ is the vector of the kth user's data. The multi-rate nature is revealed by the fact that there are many more columns, which we call signal vectors, in $A_k(\tau_k)$ which depend on the parameter of interest than in the single-rate case. Most of the time, these signal vectors are not mutually orthogonal. For more details of the system model see [6, 10].

3 Subspace Based Estimation

In this section, we will present a subspace-based algorithm for the general model described by Equations (1) and (2). This algorithm has a form which clearly exploits the correlation among signal vectors and hence is suitable for delay estimation for multi-rate systems.

3.1 Asymptotic Subspace Maximum Likelihood (ASML)

The sample data correlation matrix is given by

$$\hat{R} = \frac{1}{M}\sum_{k=1}^{M} \underline{r}(k)\underline{r}(k)^H = \left[\begin{array}{cc} \hat{E}_s & \hat{E}_n \end{array} \right] \left[\begin{array}{cc} \hat{\Lambda}_s & 0 \\ 0 & \hat{\Lambda}_n \end{array} \right] \left[\begin{array}{c} \hat{E}_s^H \\ \hat{E}_n^H \end{array} \right], \tag{3}$$

where M is the number of samples. This matrix $\hat{R}$ is the maximum likelihood estimate of the true correlation matrix R and has the eigenvalue decomposition (EVD) shown above. The true signal and noise subspaces are spanned by $E_s = \left[\begin{array}{cccc} \underline{e}_{s,1} & \underline{e}_{s,2} & \cdots & \underline{e}_{s,N_s} \end{array} \right]$ and $E_n = \left[\begin{array}{cccc} \underline{e}_{n,1} & \underline{e}_{n,2} & \cdots & \underline{e}_{n,N_g} \end{array} \right]$, respectively, where $e_{s,i}$ and $e_{n,i}$ are eigenvectors of R. The matrices Λ_s and Λ_n are diagonal, with the eigenvalues of R corresponding to signal and noise subspaces, respectively, on the diagonals. The sample version of E_s, $e_{s,i}$, E_n, $e_{n,i}$, Λ_s, and Λ_n with respect to $\hat{R}$ are denoted by $\hat{E}_s$, $\hat{e}_{s,i}$, $\hat{E}_n$, $\hat{e}_{n,i}$, $\hat{\Lambda}_s$, and $\hat{\Lambda}_n$. The dimensions of the signal subspace and noise subspace are N_s and N_g. We denote the noise variance by σ^2; thus $\Lambda_n = \sigma^2 I_{N_g}$. Assume τ_k is the delay of the user of interest.

Our objective is to determine the maximum likelihood estimate of the delay given the noise subspace. In order to specify this estimator, we define [12]

$$U = \sum_{k=1}^{N_s} \frac{\lambda_{s,k}\sigma^2}{(\lambda_{s,k}-\sigma^2)^2} \underline{e}_{s,k}\underline{e}_{s,k}^H \tag{4}$$

where $\lambda_{s,k}$ is the eigenvalue of the true covariance matrix corresponding to $\underline{e}_{s,k}$. Let P_s be the projection operator onto the signal subspace. Then, from [12], the projections of eigenvectors of sample noise subspace have asymptotically zero-mean Gaussian distribution. As $M \to \infty$,

$$P_s\hat{\underline{e}}_{n,j} \sim N(0, U/M) \tag{5}$$

$$\mathbf{E}\{(P_s\hat{\underline{e}}_{n,j})(P_s\hat{\underline{e}}_{n,l})^H\} = \frac{1}{M}U\delta_{j,l}\ , \tag{6}$$

where $\delta_{j,l} = 1$ for $j = l$ and $\delta_{j,l} = 0$ if $j \neq l$. An asymptotic subspace maximum likelihood (ASML) estimator based on the maximum likelihood of projection of signal vectors onto sample noise subspace is then given by,

$$\begin{aligned} \hat{\tau}_k &= \arg\min_{\tau_k} \frac{N_g}{M} \ln |A_k^H(\tau_k)UA_k(\tau_k)| + \\ &+ \ Tr\left\{\left[A_k^H(\tau_k)UA_k(\tau_k)\right]^{-1} A_k^H(\tau_k)\hat{E}_n\hat{E}_n^H A_k(\tau_k)\right\}. \end{aligned}$$

For M sufficiently large, the effect of the first term is negligible. Ignoring this first term, the estimator falls into the same general form as the algorithms in [6, 10] conform to

$$\hat{\tau}_k = arg\ \min_{\tau_k}\ Tr\left\{W_k A_k^H(\tau_k)\hat{E}_n\hat{E}_n^H A_k(\tau_k)\right\} \tag{7}$$

where W_k is a weighting matrix that depends on the algorithm. In [6, 10], only diagonal weighting matrices were used. The optimal weighting matrix $\left[A_k^H(\tau_k)UA_k(\tau_k)\right]^{-1}$ captures information about the asymptotic distribution and correlation of the projection of signal vectors onto the noise subspace, hence it can exploit the multi-rate structure and provide the best asymptotic performance. A constant weighting matrix or a diagonal matrix will generally yield a simpler optimization of (7); however, this comes at the expense of performance. This performance loss can be appreciable if the various signal vectors have disparate energies. This scenario occurs in the case of VSL access. Thus, for multi-rate CDMA, the optimal weighting matrix should be employed.

Since the subspace algorithm derived above is based on the general model for multi-rate systems, it can be applied to both MC and VSL access for synchronization.

3.2 Two-stage detection-estimation scheme for VSL

In VSL delay estimation, we observe that in the direct application of the previous general algorithm, the length of the relevant signal vectors can be quite long when we form the correlation matrix for low rate users. Thus, the required eigenvalue decompositions are computationally expensive. In the VSL case, each user is assigned one spreading code regardless of its data rate, but the spreading gain of the code can be different, as shown in Figure 1(b). We shall assume a specific form for VSL spreading codes for low data rate users. Repetition codes will be employed, *i.e.*, each user is assigned a spreading code with the same spreading gain N and the kth user repeats the code assigned to it $\nu(k)$

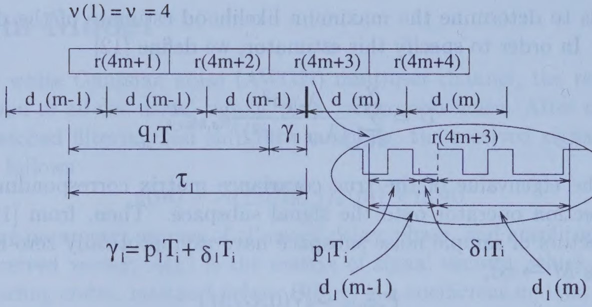


Figure 2: Two stage delay estimation in the VSL case

times to get its actual spreading code with spreading gain $\nu(k)N$. If T is the symbol duration of the highest rate users, then the symbol duration of the kth user is $\nu(k)T$. We shall propose a two-stage estimation-detection algorithm for lower-rate users, as shown in Figure 2. Assume that user 1 is the user of interest. We can express the desired user's delay as $\tau_1 = q_1 T + \gamma_1$ where $q_1 \in Z^+$ and $\gamma_1 \in [0, T)$. In the two-stage algorithm, we first estimate γ_1 using observation vectors corresponding to a duration T and then perform "location detection" to determine in which sub-interval the actual delay occurs, *i.e.*, q_1. The motivation behind this algorithm is to reduce complexity while maintaining good performance.

The first stage can be done by applying directly the proposed subspace algorithms. Let $\hat{\gamma}_1$ be the first stage estimate. Let T_1 be the symbol duration of user 1; thus $T_1 = \nu(1)T$ where $\nu(1) \in Z^+$. Location estimation is conducted by performing a $\nu(1)$-ary hypothesis test. The Bayes optimal test for equal priors and uniform costs is then,

$$\hat{q}_1 = \arg\max \Gamma_{q_1} \quad q_1 \in \{1, 2, \ldots, \nu(1)\} \tag{8}$$

where the test statistic Γ_i is given by

$$\Gamma_i = \underline{\alpha}_2^H \hat{R}_i^{-1} \underline{\alpha}_2 - \underline{\alpha}_1^H \hat{R}_i^{-1} \underline{\alpha}_1 \tag{9}$$

after some approximations. The definition of $\underline{\alpha}_1$ and $\underline{\alpha}_2$ is similar to the definition in [6]. The signal vector $\underline{\alpha}_1$ corresponds to the scenario where the current bit is the same as the previous one, and $\underline{\alpha}_2$ corresponds to the case where the two bits are different. Note that both $\underline{\alpha}_1$ and $\underline{\alpha}_2$ correspond to the duration of a sub-interval, *i.e.*, T. The matrix $\hat{R}_i$ is the correlation matrix of the received vectors which fall into the ith sub-interval. The final delay estimate is then

$$\hat{\tau}_1 = \hat{q}_1 T + \hat{\gamma}_1 \tag{10}$$

4 Performance Analysis

Our performance analysis evaluates the mean-squared error of the subspace algorithms. The asymptotic Cramér-Rao lower bound of the time delay estimator will also be considered.

4.1 Performance of Subspace Based Estimation

Recall that the general form of subspace-based estimation is

$$\hat{\tau} = \arg\min f(\tau) = \arg\min \; Tr\left\{W A^H \hat{E}_n \hat{E}_n^H A\right\} \tag{11}$$

with the subscript k suppressed. An exact analysis is challenging due to the need to manipulate the true distribution of $\hat{E}_n$. Thus, we investigate the asymptotic performance of this type of algorithm. The estimate $\hat{\tau}$ is a minimum point of $f(\tau)$. Thus we have, using Taylor expansion of $f(\tau)$ around the true value of τ,

$$\frac{df(\hat{\tau})}{d\hat{\tau}} \simeq \frac{df(\tau)}{d\tau} + \frac{d^2 f(\tau)}{d\tau^2}(\hat{\tau} - \tau) = 0 \,. \tag{12}$$

That implies

$$\hat{\tau} - \tau \simeq -\frac{df(\tau)/d\tau}{d^2 f(\tau)/d\tau^2} \,. \tag{13}$$

So, we get the asymptotic variance of $\hat{\tau} - \tau$

$$\mathbf{E}\left\{(\hat{\tau} - \tau)(\hat{\tau} - \tau)^*\right\} = \mathbf{E}\left\{\frac{|df(\tau)/d\tau|^2}{|d^2 f(\tau)/d\tau^2|^2}\right\} \,. \tag{14}$$

From the appendices in [12], we know $E_s^H \hat{E}_n$ and $-\hat{E}_s^H E_n$ have the same asymptotic distribution. And

$$A^H \hat{E}_n \hat{E}_n^H \simeq -A^H E_s (\hat{E}_s^H E_n) E_n^H \,. \tag{15}$$

Let $E_s E_s^H \hat{E}_n = \tilde{E}_n = \left[\begin{array}{cccc} \tilde{e}_{n,1} & \tilde{e}_{n,2} & \cdots & \tilde{e}_{n,N_g} \end{array}\right]$, then

$$\mathbf{E}\left\{\tilde{e}_{n,i} \tilde{e}_{n,j}^H\right\} = \frac{1}{M} U \delta_{i,j} \tag{16}$$

where U is defined in (4). So, as $M \to \infty$, we get the expression of asymptotic variance of $\hat{\tau} - \tau$ as

$$\mathbf{E}\left\{(\hat{\tau} - \tau)(\hat{\tau} - \tau)^*\right\} = \frac{Tr\left\{E_n^H \Psi W^H A^H U A W \Psi^H E_n\right\}}{2MTr^2\left\{W \Psi^H E_n E_n^H \Psi\right\}} \tag{17}$$

The matrix $\Psi = \frac{dA}{d\tau}$ is the derivative of the signal matrix with respect to τ.

4.2 Performance analysis for the location detection

The performance of the location detector in (9) is evaluated via determination of the probability of selecting the incorrect sub-interval. We assume that the data samples are mutually independent, then, the sample covariance matrix $M\hat{R}$ where M is the number of samples can be shown to have a central Wishart distribution with M degrees of freedom. Due to the sample independence assumption, the test statistics are also independent of each other. Let $R_i = FF^H$, thus $\hat{R}_i$ can be written into

$$\hat{R}_i = F\left(I + \Delta\right) F^H \tag{18}$$

The components of $\hat{R}_i$, hence Δ, are shown to be asymptotically Gaussian [13]. As M becomes sufficiently large, we can approximate $\hat{R}_i^{-1}$ by

$$\hat{R}_i^{-1} = F^{-H}\left(I + \Delta\right)^{-1} F^{-1} \approx R_i^{-1} - F^{-H} \Delta F^{-1} \tag{19}$$

Hence, the decision metrics Γ_i can be approximated by a Gaussian random variable. Based on the results in [13], the mean and variance of the test statistics Γ_i is given by:

$$\zeta_i = \mathbf{E}\left\{\Gamma_i\right\} = \frac{M}{M-p}\left(\underline{\alpha}_2^H R_i^{-1} \underline{\alpha}_2 - \underline{\alpha}_1^H R_i^{-1} \underline{\alpha}_1\right) \tag{20}$$

$$\begin{aligned} \sigma_i^2 &= \mathbf{Var}\left\{\Gamma_i\right\} = \frac{M^2\left(\underline{\alpha}_2^H R_i^{-1} \underline{\alpha}_2 - \underline{\alpha}_1^H R_i^{-1} \underline{\alpha}_1\right)^2}{(M-p+1)(M-p)^2(M-p-1)} \\ &+ \frac{M^2\left[\left(\underline{\alpha}_2^H R_i^{-1} \underline{\alpha}_2\right)^2 - 2\left|\underline{\alpha}_2^H R_i^{-1} \underline{\alpha}_1\right|^2 + \left(\underline{\alpha}_1^H R_i^{-1} \underline{\alpha}_1\right)^2\right]}{(M-p+1)(M-p)(M-p-1)} \end{aligned}$$

where p is the length of the sample vectors and R_i is the correlation matrix of the received vectors in the ith sub-interval. Thus, the detection error probability (DER) when k is the correct sub-interval can be approximated by its union bound, which is given by

$$P_e \approx 1 - \prod_{i=1, i \neq k}^{\nu(1)} \left\{1 - Q\left(\frac{|\zeta_k - \zeta_i|}{\sqrt{\sigma_k^2 + \sigma_i^2}}\right)\right\} \tag{21}$$

5 Numerical results

In this section, we present numerical examples which enable comparison of the performance of the proposed synchronization methods. Gold sequences with length $N = 31$ [14] generated by the characteristic polynomials $g_1(p) = p^5 + p^2 + 1$ and $g_2(p) = p^5 + p^4 + p^2 + p + 1$ are employed as user spreading codes. Delays are generated via realizations of mutually independent random variables uniformly distributed over the range they could occur. In the MC case, the range of the delays is $[0, T)$ whereas in the VSL case, the range of the delay of the kth user is $[0, \nu(k)T)$.

In the MC case, the simulated system is a three-rate system with one user with four times the base rate; one user with two times the base rate; and four base rate users. In Figure 3, the performance of the proposed subspace (ASML), and conventional subspace methods (SUB), are compared with the CRB for the MC case. The SUB method is a direct extension of the algorithms in [6, 9] to the multi-rate case. Thus diagonal weighting matrices are employed. The ASML algorithm has lower mean squared error (MSE) than the SUB algorithm as expected. In this case, these two algorithms perform comparably. For both ASML and conventional subspace methods, the simulation results match the analyses quite well. The complexity of the ASML algorithm depends on the numerical techniques used to search for the desired minima.

For VSL access, the performance of ASML and SUB is compared with the CRB for the VSL case, as shown in Figure 4. The system is a dual-rate, ten-user system with five high rate users and five low rate users. The rate ratio is $\eta = 4$. We focus primarily on estimation algorithms over the short interval (the symbol interval of high rate users) with the understanding that a two-stage algorithm will be used. As M goes large, ASML performs much better than the conventional subspace method on the same period. The use of the two-stage strategy reduces complexity (in this case) by a factor of $\eta^3 = 64$.

The performance of the location detection algorithm used in VSL is investigated in Figure 5. The same system as in Figure 4 is considered. It is observed that the detection algorithm can achieve a small detection error probability with a limited number of data samples and mild SNR conditions. The simulation results matches the analytical results reasonably well, but they do not match perfectly due to the assumption that all data samples are independent in the performance analysis. In practice, this assumption does not hold. However, it is anticipated that one should be able to incorporate the effects of correlation in a future analysis.

6 Conclusion

In this paper, we have considered the problem of delay estimation for the multi-rate asynchronous DS-CDMA systems with multi-code or variable spreading length multi-rate access schemes. General results on maximum likelihood based subspace time delay estimation were provided. A contribution of the paper is the derivation of the asymptotically optimal weighting matrix. For single-rate time delay estimation, suboptimal diagonal approximation of this weighting matrix had previously been considered. However, the results herein show that the optimal weighting matrix can offer significant performance improvement for the multi-rate scenarios, especially in the case of VSL access. To reduce complexity, a two-stage estimator is developed for VSL access. It is observed that the reduction in complexity does not come at the expense of performance. Asymptotic approximations of the mean-squared error are derived, which appear to be tight for moderate SNR's and a moderate number of observations. Finally, the detection performance of the two-stage algorithm is analytically studied.

References

[1] Urbashi Mitra. Comparison of Maximum-Likelihood-Based Detection for Two Multi-rate Access Schemes for CDMA Signals. *IEEE Trans. Comm.*, 47(1):64–77, January 1999.

[2] T. Ottosson and A. Svensson. Multi-Rate Schemes in DS/CDMA Systems. In *Proc. 1995 IEEE 45th Veh. Tech. Conf.*, pages 1006–1010, Chicago, IL, July 1995.

[3] A. Sabharwal, Urbashi Mitra, and R. Moses. Cyclic Wiener Filtering Based Multi-rate DS-CDMA Receivers. In *Proceedings of IEEE WCNC'99*, New Orleans, September 1999.

[4] U. Madhow and M. L. Honig. MMSE interference supression for timing acquisition and demodulation in direct-sequence CDMA systems. *IEEE Trans. Comm.*, 46(8):1065–1075, August 1998.

[5] S. E. Bensley and B. Aazhang. Maximum-Likelihood Synchronization of a Single User for Code-Division Multiple-Access Communication Systems. *IEEE Trans. Comm.*, 46(3):392–399, March 1998.

[6] E. G. Ström, S. Parkvall, S. L. Miller, and B. E. Ottersten. Propagatin Delay Estimation in Asynchronous Direct-Sequence Code-Division Multiple Access Systems. *IEEE Trans. Comm.*, 44(1):84–93, January 1996.

[7] S. E. Bensley and B. Aazhang. Subspace-Based Channel Estimation for Code Division Multiple Access Communication Systems. *IEEE Trans. Comm.*, 44(8):1009–1020, August 1996.

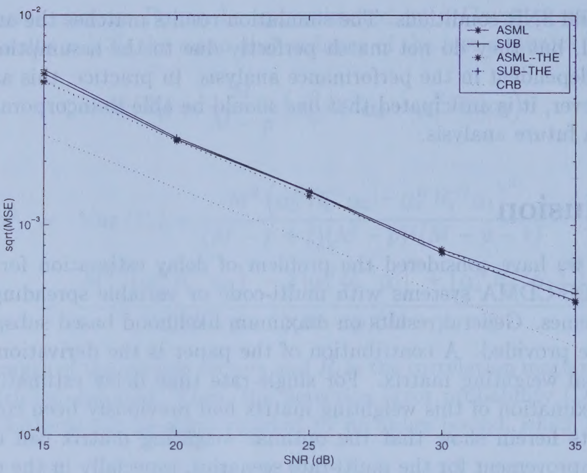


Figure 3: Mean squared error for the proposed subspace algorithms in the MC case

[8] M. Torlak and G. Xu. Blind Multiuser Channel Estimation in Asynchronous CDMA Systems. *IEEE Trans. Signal Processing*, 45(1):137–147, January 1997.

[9] T. Östman, S. Parkvall, and B. E. Otternsten. An Improved MUSIC Algorithm for Estimation of Time Delays in Asynchronous DS-CDMA Systems. *IEEE Trans. Comm.*, 47(11):1628–1631, November 1999.

[10] T. Östman, B. E. Ottersten, and L. B. Milstein. Estimation Of Time Delays In Multirate Asynchronous DS-CDMA Systems. In *Proceedings of ISSSTA 98*, Sun City, South Africa, September 1998.

[11] S. Parkvall. *Near-far resistant DS-CDMA systems: Parameter estimation and data detection.* PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1996.

[12] P. Stoica and A. Nehorai. MUSIC, Maximum Likelihood, and Cramér-Rao Bound. *IEEE Trans. Acoust., Speech, Signal Processing*, 37(5):720–741, May 1989.

[13] John A. Tague and Curtis I. Caldwell. Expectations of Useful Complex Wishart Forms. *Multidimensional Systems and Signal Processing*, 5(3):263–279, July 1994.

[14] John G. Proakis. *Digital Communications.* McGraw-Hill Book Company, 1995.

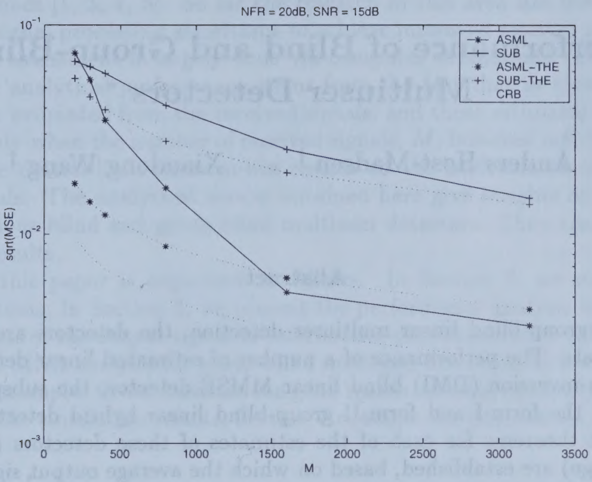


Figure 4: Mean squared error for the proposed subspace algorithms in the VSL case

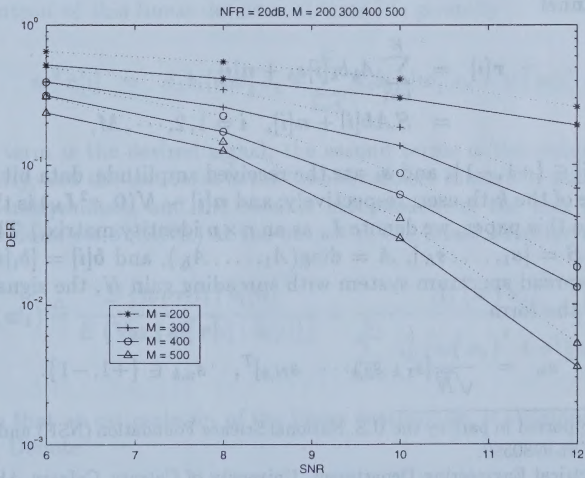


Figure 5: Performance of location detection algorithm in the VSL case (NFR = 20dB).

Performance of Blind and Group-Blind Multiuser Detectors *

Anders Høst-Madsen † Xiaodong Wang ‡

Abstract

In blind and group-blind linear multiuser detection, the detectors are estimated from the received signals. The performance of a number of estimated linear detectors, including the direct-matrix-inversion (DMI) blind linear MMSE detector, the subspace blind linear MMSE detector, the form-I and form-II group-blind linear hybrid detectors, is analyzed. Asymptotic limit theorems for each of the estimates of these detectors (when the signal sample size is large) are established, based on which the average output signal-to-inference-plus-noise ratios (SINR's) of these detectors are given. The analytical SINR expressions obtained here of various estimated linear detectors match well with the simulation results.

1 Introduction

This paper is concerned with the analysis of the performance of blind and group-blind linear multiuser detection techniques for the basic discrete-time synchronous CDMA multiple-access K-user channel

$$\begin{aligned} \boldsymbol{r}[i] &= \sum_{k=1}^{K} A_k b_k[i] \boldsymbol{s}_k + \boldsymbol{n}[i] && (1) \\ &= \boldsymbol{S}\boldsymbol{A}\boldsymbol{b}[i] + \boldsymbol{n}[i], \quad i = 1, 2, \cdots, M, && (2) \end{aligned}$$

where $A_k > 0$, $b_k[i] \in \{+1, -1\}$, and $\boldsymbol{s}_k$ are the received amplitude, data bit, and unit-energy signature sequence of the k-th user, respectively; and $\boldsymbol{n}[i] \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$ is the additive white Gaussian noise. (In this paper, we denote $\boldsymbol{I}_n$ as an $n \times n$ identity matrix.) These are collected in vector form as $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K]$, $\boldsymbol{A} = \text{diag}(A_1, \ldots, A_K)$, and $\boldsymbol{b}[i] = [b_1[i], \ldots, b_K[i]]^T$. In a direct-sequence spread spectrum system with spreading gain N, the signature sequence of the k-th user is of the form

$$\boldsymbol{s}_k = \frac{1}{\sqrt{N}} [s_{1,k}\ s_{2,k}\ \cdots\ s_{N,k}]^T, \quad s_{n,k} \in \{+1, -1\}. \tag{3}$$

---

[*] This work was supported in part by the U.S. National Science Foundation (NSF) under grants CAREER CCR-9875314 and CCR-9980599.

[†] TRLabs and Electrical Engineering Department, University of Calgary, Calgary, Alberta, Canada T2L 2K7.

[‡] Electrical Engineering Department, Texas A&M University, College Station, TX 77843-3128, USA.

This work is motivated by the recent development of blind and group-blind multiuser detection techniques [1, 3, 4, 5]. So far the research in this area has been focused on the development of signal processing algorithms to achieve improved receiver performance. And the performance assessment is largely done via computer simulations. The main difficulty in obtaining the analytical performance stems from the fact that in these blind methods, the detectors are estimated from the received signals; and those estimates coincide with the true detectors only when the number of received signals, M, becomes infinitely large. In this paper, we derive central-limit-theorem-like results on the performance of these multiuser detection methods. The analytical results obtained here give insights on the performance difference of various blind and group-blind multiuser detectors. They also match well with the simulation results.

The rest of this paper is organized as follows. In Section 2, we state the problems under considerations; In Section 3, we present the performance analysis for blind multiuser detectors, in terms of the output signal-to-interference-plus-noise ratios (SINR's); In Section 4, we present the SINR analysis for group-blind multiuser detectors; In Section 5, we present some numerical examples to demonstrate the good match between our analytical performance assessment and the simulation results; Section 7 contains the conclusions.

2 Problem Statement

Consider the signal model (1). Throughout this paper, we assume that user 1 is the user of interest. A linear detector for user 1 is a (deterministic) vector $\boldsymbol{w}_1 \in \mathbb{R}^N$ such that $b_1[i]$ is demodulated according to

$$\hat{b}_1[i] \;=\; \operatorname{sign}\left(\boldsymbol{w}_1^T \boldsymbol{r}[i]\right). \tag{4}$$

Using (1), the output of this linear detector at time i is given by

$$\boldsymbol{w}_1^T \boldsymbol{r}[i] \;=\; A_1 b_1[i] \boldsymbol{w}_1^T \boldsymbol{s}_1 + \sum_{k=2}^{K} A_k b_k[i] \boldsymbol{w}_1^T \boldsymbol{s}_k + \boldsymbol{w}_1^T \boldsymbol{n}[i]. \tag{5}$$

In (5), the first term is the desired signal, the second terms is the residual multiple-access interference (MAI), and the last term is the ambient noise. Since it is assumed that the user bit streams are independent, and the noise is independent of the user bits, the signal-to-interference-plus-noise ratio (SINR) at the output of the linear detector $\boldsymbol{w}_1$ is given by

$$\operatorname{SINR}(\boldsymbol{w}_1) \triangleq \frac{E\left\{\boldsymbol{w}_1^T \boldsymbol{r}[i] \mid b_1[i]\right\}^2}{E\left\{\operatorname{Var}\left\{\boldsymbol{w}_1^T \boldsymbol{r}[i] \mid b_1[i]\right\}\right\}} = \frac{A_1^2\left(\boldsymbol{w}_1^T \boldsymbol{s}_1\right)^2}{\sum_{k=2}^{K} A_k^2\left(\boldsymbol{w}_1^T \boldsymbol{s}_k\right)^2 + \sigma^2 \|\boldsymbol{w}_1\|^2}. \tag{6}$$

Now suppose that an estimate $\hat{\boldsymbol{w}}_1$ of the linear detector $\boldsymbol{w}_1$ is obtained from the received signal $\{\boldsymbol{r}[i]\}_{i=1}^{M}$. Denote

$$\Delta \boldsymbol{w}_1 \;\triangleq\; \hat{\boldsymbol{w}}_1 - \boldsymbol{w}_1. \tag{7}$$

Apparently both $\hat{\boldsymbol{w}}_1$ and $\Delta\boldsymbol{w}_1$ are random vectors and are functions of the random quantities $\{\boldsymbol{b}[i], \boldsymbol{n}[i]\}_{i=1}^{M}$. The output of the estimated linear detector $\hat{\boldsymbol{w}}_1$ at time i is given by

$$\hat{\boldsymbol{w}}_1^T \boldsymbol{r}[i] \;=\; \boldsymbol{w}_1^T \boldsymbol{r}[i] + \Delta\boldsymbol{w}_1^T \boldsymbol{r}[i], \tag{8}$$

where the first term in (8) represents the output of the true linear detector $\boldsymbol{w}_1$, given by (5). The second term in (8) represents an additional noise term caused by the estimation error $\Delta\boldsymbol{w}_1$. Since $\Delta\boldsymbol{w}_1$ is a function of $\{\boldsymbol{r}[i]\}_{i=1}^{M}$, $\Delta\boldsymbol{w}_1$ and $\boldsymbol{r}[i]$ are correlated for a fixed i. For large M, such a correlation is negligible. In what follows we assume that $\Delta\boldsymbol{w}_1$ and $\boldsymbol{r}[i]$ are uncorrelated. Denote $\boldsymbol{C}_w \triangleq E\left\{\Delta\boldsymbol{w}_1 \Delta\boldsymbol{w}_1^T\right\}$. Then the average SINR at the output of the estimated linear detector $\hat{\boldsymbol{w}}_1$ is given by

$$\overline{\mathrm{SINR}(\hat{\boldsymbol{w}}_1)} = \frac{A_1^2 \left(\boldsymbol{w}_1^T \boldsymbol{s}_1\right)^2}{\sum_{k=2}^{K} A_k^2 \left(\boldsymbol{w}_1^T \boldsymbol{s}_k\right)^2 + \sigma^2 \|\boldsymbol{w}_1\|^2 + \mathrm{tr}\left(\boldsymbol{C}_w \boldsymbol{C}_r\right)}. \tag{9}$$

In this paper, we analyze the average output SINR of the following estimated linear detectors

- Direct-matrix-inversion (DMI) blind linear MMSE detector [1, 2];
- Subspace blind linear MMSE detector [5];
- Subspace group-blind linear hybrid detectors (form-I and form-II) [3].

The first two of the above detectors can be estimated directly based on the received signals $\{\boldsymbol{r}[i]\}_{i=1}^{M}$, with the only prior knowledge of the spreading sequence $\boldsymbol{s}_1$ of the desired user. For the group-blind detectors, it is assumed that in addition to $\boldsymbol{s}_1$, the spreading sequences of some other users, e.g., users within a cell, are known to the receiver.

We first establish asymptotic limit theorems for each of the estimates of the above detectors. That is, we show that *for fixed K and N, and fixed set of spreading sequences $\{\boldsymbol{s}_k\}_{k=1}^{K}$,*

$$\sqrt{M}\left(\hat{\boldsymbol{w}}_1 - \boldsymbol{w}_1\right) \;\to\; \mathcal{N}(\boldsymbol{0},\, M \cdot \boldsymbol{C}_w), \quad \text{in distribution, as } M \to \infty.$$

The limit error covariance $\boldsymbol{C}_w$ is given for each of the above detectors, based on which the average output SINR, $\overline{\mathrm{SINR}(\hat{\boldsymbol{w}}_1)}$, is computed. As will be seen, the analytical SINR expressions obtained in this paper of various estimated linear detectors match well with the simulations results even for finite sample size M.

3 Asymptotic Output SINR for Blind Detectors

Consider again the signal model (1). The linear MMSE detector for user 1 is defined as

$$\boldsymbol{w}_1 \;=\; \min_{\boldsymbol{w} \in \mathrm{R}^N} E\left\{\left(\frac{1}{\mu} b_1[i] - \boldsymbol{w}^T \boldsymbol{r}[i]\right)^2\right\} \;=\; \frac{A_1}{\mu} \boldsymbol{C}_r^{-1} \boldsymbol{s}_1, \tag{10}$$

with

$$C_r \triangleq E\left\{r[i]r[i]^T\right\} = \sum_{k=1}^{K} A_k^2 s_k s_k^T + \sigma^2 I_N = SA^2S^T + \sigma^2 I_N, \tag{11}$$

where μ is some positive constant. Since the linear detection rule (4) is invariant to a positive scaling, the linear detector in (10) is invariant to the positive constant A_1/μ. For simplicity we choose $\mu = A_1$ so that $w_1 = C_r^{-1}s_1$.

Let the eigendecomposition of C_r in (11) be

$$C_r \;=\; U_s\Lambda_sU_s^T + \sigma^2U_nU_n^T, \tag{12}$$

where $\Lambda_s = \mathsf{diag}(\lambda_1, \cdots, \lambda_K)$ contains the largest K eigenvalues of C_r; $U_s = [u_1, \cdots, u_K]$ contains the eigenvectors corresponding to the largest K eigenvalues in Λ_s; $U_n = [u_{K+1}, \cdots, u_N$ contains the $(N-K)$ eigenvectors corresponding to the smallest eigenvalue σ^2 of C_r. It is known that $\mathsf{range}(U_s) = \mathsf{range}(S)$ is the signal subspace; and $\mathsf{range}(U_n) \perp \mathsf{range}(S)$ is the noise subspace. The linear MMSE detector w_1 in (10) can also be written in terms of the signal subspace components as [5]

$$w_1 \;=\; U_s\Lambda_s^{-1}U_s^Ts_1. \tag{13}$$

Corresponding to the two forms of the linear MMSE detector (10) and (13), there are two approaches to its blind implementation. In the direct-matrix-inversion (DMI) method, the autocorrelation matrix C_r in (10) is replaced by the corresponding sample estimate. That is

$$\hat{C}_r \;=\; \frac{1}{M}\sum_{i=1}^{M} r[i]r[i]^T, \tag{14}$$

$$\hat{w}_1 \;=\; \hat{C}_r^{-1}s_1. \qquad \text{[DMI blind linear MMSE detector]} \tag{15}$$

In the subspace method, the eigencomponents Λ_s and U_s in (13) are replaced by the corresponding eigenvalues and eigenvectors of the sample autocorrelation matrix $\hat{C}_r$. That is,

$$\begin{aligned} \hat{C}_r &= \frac{1}{M}\sum_{i=1}^{M} r[i]r[i]^T \\ &= \hat{U}_s\hat{\Lambda}_s\hat{U}_s^T + \hat{U}_n\hat{\Lambda}_n\hat{U}_n^T, && (16) \\ \hat{w}_1 &= \hat{U}_s\hat{\Lambda}_s^{-1}\hat{U}_s^Ts_1, \qquad \text{[Subspace blind linear MMSE detector]} && (17) \end{aligned}$$

where $\hat{\Lambda}_s$ and $\hat{U}_s$ contain respectively the largest K eigenvalues and the corresponding eigenvectors of $\hat{C}_r$; $\hat{\Lambda}_n$ and $\hat{U}_n$ contain respectively the remaining eigenvalues and eigenvectors of $\hat{C}_r$. The following result gives the asymptotic distribution of the blind linear MMSE detectors given by (15) and (17).

Theorem 1 *Let w_1 be the true linear MMSE detector given by (10) [or equivalently (13)], and $\hat{w}_1$ be the estimated blind linear MMSE detector given by (15) or (17). Let (12) be the eigendecomposition of the autocorrelation matrix C_r of the received signals. Then*

$$\sqrt{M}\,(\hat{w}_1 - w_1) \;\to\; \mathcal{N}(0, M\cdot C_w), \quad \textit{in distribution, as } M \to \infty,$$

with

$$\boldsymbol{C}_w = \frac{1}{M}\left[\left(\boldsymbol{w}_1^T\boldsymbol{s}_1\right)\boldsymbol{U}_s\boldsymbol{\Lambda}_s^{-1}\boldsymbol{U}_s^T + \boldsymbol{w}_1\boldsymbol{w}_1^T - 2\boldsymbol{U}_s\boldsymbol{\Lambda}_s^{-1}\boldsymbol{U}_s^T\boldsymbol{S}\boldsymbol{D}\boldsymbol{S}^T\boldsymbol{U}_s\boldsymbol{\Lambda}_s^{-1}\boldsymbol{U}_s^T + \tau\boldsymbol{U}_n\boldsymbol{U}_n^T\right], \quad (18)$$

where

$$\begin{aligned}\boldsymbol{D} &\triangleq \mathsf{diag}\left\{A_1^4\left(\boldsymbol{w}_1^T\boldsymbol{s}_1\right)^2, A_2^4\left(\boldsymbol{w}_1^T\boldsymbol{s}_2\right)^2, \cdots, A_K^4\left(\boldsymbol{w}_1^T\boldsymbol{s}_K\right)^2\right\}, &(19)\\ \tau &\triangleq \begin{cases}\frac{1}{\sigma^2}\boldsymbol{s}_1^T\boldsymbol{U}_s\boldsymbol{\Lambda}_s^{-1}\boldsymbol{U}_s^T\boldsymbol{s}_1, & \textit{DMI blind detector}\\ \sigma^2\boldsymbol{s}_1^T\boldsymbol{U}_s\boldsymbol{\Lambda}_s^{-1}\left(\boldsymbol{\Lambda}_s - \sigma^2\boldsymbol{I}_K\right)^{-2}\boldsymbol{U}_s^T\boldsymbol{s}_1, & \textit{subspace blind detector}\end{cases}. &(20)\end{aligned}$$

Hence for large M, the covariance of the blind linear detector, $\boldsymbol{C}_w = E\left\{\Delta\boldsymbol{w}_1\Delta\boldsymbol{w}_1^T\right\}$, can be approximated by (18). Denote

$$\boldsymbol{R} \triangleq \boldsymbol{S}^T\boldsymbol{S}. \quad (21)$$

The next result gives the expression for the average output SINR, defined by (9), of the blind linear detectors.

Corollary 1 *The average output SINR of the estimated blind linear detector is given by* $\overline{\mathrm{SINR}(\hat{\boldsymbol{w}}_1)}$

$$= \frac{A_1^2\left(\boldsymbol{w}_1^T\boldsymbol{s}_1\right)^2}{\sum_{k=2}^{K}A_k^2\left(\boldsymbol{w}_1^T\boldsymbol{s}_k\right)^2 + \sigma^2\|\boldsymbol{w}_1\|^2 + \frac{1}{M}\left[(K+1)\boldsymbol{w}_1^T\boldsymbol{s}_1 - 2\sum_{k=1}^{K}A_k^4\left(\boldsymbol{w}_1^T\boldsymbol{s}_k\right)^2\left(\boldsymbol{w}_k^T\boldsymbol{s}_k\right) + (N-K)\tau\sigma^2\right]}, \quad (22)$$

where

$$\begin{aligned}\boldsymbol{w}_l^T\boldsymbol{s}_k &= \frac{1}{A_l^2}\left[\boldsymbol{R}\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\right]_{k,l}, \quad k,l = 1,\cdots,K, &(23)\\ \|\boldsymbol{w}_1\|^2 &= \frac{1}{A_1^4}\left[\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\right]_{1,1}, &(24)\\ \tau\sigma^2 &= \begin{cases}\boldsymbol{w}_1^T\boldsymbol{s}_1, & \textit{DMI blind detector}\\ \frac{\sigma^4}{A_1^4}\left[\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\boldsymbol{A}^{-2}\boldsymbol{R}^{-1}\right]_{1,1}, & \textit{subspace blind detector}\end{cases}. &(25)\end{aligned}$$

It is seen from (22) that the performance difference between the DMI blind detector and the subspace blind detector is caused by the single parameter τ given by (20) - the detector with a smaller τ has a higher output SINR.

4 Asymptotic Output SINR for Group-blind Detectors

In this section, we consider the performance of group-blind linear multiuser detection. Specifically, we focus on the performance of the group-blind linear hybrid detectors [3]. Recall that in blind linear multiuser detection, it is assumed that the receiver has the knowledge of only the spreading sequence of the desired user. In group-blind multiuser detection, however, it is assumed that the receiver has the knowledge of the spreading sequences of some but not all the users. Without loss of generality, assume that the first $\tilde{K}$ users' spreading sequences are known to the receiver, whereas those of the rest $(K - \tilde{K})$ users' are unknown. Denote $\tilde{\boldsymbol{S}} \triangleq [\boldsymbol{s}_1 \cdots \boldsymbol{s}_{\tilde{K}}]$. It is assumed that $\tilde{\boldsymbol{S}}$ has full column rank. Denote $\tilde{\boldsymbol{e}}_k$ as the k-th unit vector in $\mathbb{R}^{\tilde{K}}$. The group-blind linear hybrid detector zero-forces the interference caused by the $\tilde{K}$ known users and suppress that from the rest $(K - \tilde{K})$ unknown users according to the MMSE criterion. In particular, such a detector for user 1 is given by the solution to the following constrained optimization problem

$$\boldsymbol{w}_1 = \arg \min_{\boldsymbol{w} \in \mathbb{R}^N} E\left\{ \left(A_1 b_1[i] - \boldsymbol{w}^T \boldsymbol{r}[i] \right)^2 \right\}, \qquad \text{s.t. } \boldsymbol{w}^T \tilde{\boldsymbol{S}} = \tilde{\boldsymbol{e}}_1^T. \tag{26}$$

In [3], two forms of group-blind hybrid detectors are given. In what follows, we find the asymptotic distributions of both forms of the group-blind hybrid detectors.

4.1 Form-II Group-blind Hybrid Detector

It is shown in [3] that the solution to the constrained optimization problem (26) is given in terms of the signal subspace components of $\boldsymbol{C}_r$ [cf.(12)] as

$$\boldsymbol{w}_1 = \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \tilde{\boldsymbol{S}} \left(\tilde{\boldsymbol{S}}^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \tilde{\boldsymbol{S}} \right)^{-1} \tilde{\boldsymbol{e}}_1. \tag{27}$$

The form-II group-blind linear hybrid detector is formed by replacing the eigencomponents of $\boldsymbol{C}_r$ in (27) by those of the corresponding sample correlation matrix $\hat{\boldsymbol{C}}_r$.

The following result gives the asymptotic distribution of the form-II group-blind hybrid detector.

Theorem 2 *Let $\hat{\boldsymbol{w}}_1$ be the estimated form-II group-blind linear hybrid detector, given by*

$$\hat{\boldsymbol{w}}_1 = \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s^{-1} \hat{\boldsymbol{U}}_s^T \tilde{\boldsymbol{S}} \left(\tilde{\boldsymbol{S}}^T \hat{\boldsymbol{U}}_s \hat{\boldsymbol{\Lambda}}_s^{-1} \hat{\boldsymbol{U}}_s^T \tilde{\boldsymbol{S}} \right)^{-1} \tilde{\boldsymbol{e}}_1. \tag{28}$$

Then

$$\sqrt{M}\,(\hat{\boldsymbol{w}}_1 - \boldsymbol{w}_1) \;\to\; \mathcal{N}(\boldsymbol{0}, M \cdot \boldsymbol{C}_w), \quad \text{in distribution, as } M \to \infty,$$

with

$$\boldsymbol{C}_w = \frac{1}{M} \left\{ \boldsymbol{Q} \left[\left(\boldsymbol{w}_1^T \boldsymbol{v}_1 \right) \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T - 2 \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \boldsymbol{S} \boldsymbol{D} \boldsymbol{S}^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \right] \boldsymbol{Q}^T + \tau \boldsymbol{U}_n \boldsymbol{U}_n^T \right\}, \tag{29}$$

where

$$\boldsymbol{v}_1 \triangleq \tilde{\boldsymbol{S}} \left(\tilde{\boldsymbol{S}}^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \tilde{\boldsymbol{S}} \right)^{-1} \tilde{\boldsymbol{e}}_1, \tag{30}$$

$$\boldsymbol{D} \triangleq \mathsf{diag}\left\{ A_1^4 \left(\boldsymbol{w}_1^T \boldsymbol{s}_1 \right)^2, \cdots, A_K^4 \left(\boldsymbol{w}_1^T \boldsymbol{s}_K \right)^2 \right\}, \tag{31}$$

$$\tau = \sigma^2 \boldsymbol{v}_1^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \left(\boldsymbol{\Lambda}_s - \sigma^2 \boldsymbol{I}_K \right)^{-2} \boldsymbol{U}_s^T \boldsymbol{v}_1, \tag{32}$$

$$\boldsymbol{Q} \triangleq \boldsymbol{I}_N - \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \tilde{\boldsymbol{S}} \left(\tilde{\boldsymbol{S}}^T \boldsymbol{U}_s \boldsymbol{\Lambda}_s^{-1} \boldsymbol{U}_s^T \tilde{\boldsymbol{S}} \right)^{-1} \tilde{\boldsymbol{S}}^T. \tag{33}$$

Define the partition of the following matrix

$$\boldsymbol{R}\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\boldsymbol{A}^{-2} = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{12}^T & \boldsymbol{\Psi}_{22} \end{bmatrix}, \tag{34}$$

where the dimension of $\boldsymbol{\Psi}_{11}$ is $\tilde{K} \times \tilde{K}$. Note that the left-hand side of (34) is equal to $\left(\boldsymbol{S}^T\boldsymbol{U}_s\boldsymbol{\Lambda}_s^{-1}\boldsymbol{U}_s^T\boldsymbol{S}\right)$ and therefore it is symmetric. Define further

$$\begin{aligned} \boldsymbol{\Pi} &\triangleq \left[\boldsymbol{A}^{-2}\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\boldsymbol{A}^{-2}\right]_{1:\tilde{K},1:\tilde{K}}, &(35)\\ \boldsymbol{\Xi} &\triangleq \left[\boldsymbol{A}^{-2}\left(\boldsymbol{R}+\sigma^2\boldsymbol{A}^{-2}\right)^{-1}\boldsymbol{A}^{-2}\boldsymbol{R}^{-1}\boldsymbol{A}^{-2}\right]_{1:\tilde{K},1:\tilde{K}}. &(36)\end{aligned}$$

The next result gives the expression for the average output SINR of the form-II group-blind hybrid detector.

Corollary 2 *The average output SINR of the estimated form-II group-blind linear hybrid detector is given by*

$$\mathrm{SINR}(\hat{\boldsymbol{w}}_1) = \frac{A_1^2}{\sum_{k=1}^{K-\tilde{K}} A_{\tilde{K}+k}^2\left(\boldsymbol{w}_1^T\boldsymbol{s}_{\tilde{K}+k}\right)^2 + \sigma^2\|\boldsymbol{w}_1\|^2 + \mathrm{tr}(\boldsymbol{C}_w\boldsymbol{C}_r)}, \tag{37}$$

where

$$\begin{aligned} \boldsymbol{w}_1^T\boldsymbol{s}_{\tilde{K}+k} &= \left[\boldsymbol{\Psi}_{12}^T\boldsymbol{\Psi}_{11}^{-1}\right]_{k,1}, &(38)\\ \|\boldsymbol{w}_1\|^2 &= \left[\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Pi}\boldsymbol{\Psi}_{11}^{-1}\right]_{1,1}, &(39)\\ \mathrm{tr}(\boldsymbol{C}_w\boldsymbol{C}_r) &= \frac{1}{M}\Bigg\{(K-\tilde{K})\left[\boldsymbol{\Psi}_{11}^{-1}\right]_{1,1} - 2\sum_{k=1}^{K-\tilde{K}} A_{\tilde{K}+k}^4\left[\boldsymbol{\Psi}_{12}^T\boldsymbol{\Psi}_{11}^{-1}\right]_{k,1}^2\left[\boldsymbol{\Psi}_{22}-\boldsymbol{\Psi}_{12}^T\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Psi}_{12}\right]_{k,k} \\ &\quad +(N-K)\sigma^4\left[\boldsymbol{\Psi}_{11}^{-1}\boldsymbol{\Xi}\boldsymbol{\Psi}_{11}^{-1}\right]_{1,1}\Bigg\}. &(40)\end{aligned}$$

4.2 Form-I Group-blind Detectors

Define the following projection matrix

$$\bar{\boldsymbol{P}} \triangleq \boldsymbol{I}_N - \tilde{\boldsymbol{S}}\left(\tilde{\boldsymbol{S}}^T\tilde{\boldsymbol{S}}\right)^{-1}\tilde{\boldsymbol{S}}^T, \tag{41}$$

which projects any signal onto the orthogonal subspace null $\left(\tilde{\boldsymbol{S}}^T\right)$. It is then easily seen that the matrix $\left(\bar{\boldsymbol{P}}\boldsymbol{C}_r\bar{\boldsymbol{P}}\right)$ has an eigendecomposition of the form

$$\bar{\boldsymbol{P}}\boldsymbol{C}_r\bar{\boldsymbol{P}} = [\bar{\boldsymbol{U}}_s\,\bar{\boldsymbol{U}}_n\,\bar{\boldsymbol{U}}_o]\begin{bmatrix} \bar{\boldsymbol{\Lambda}}_s & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2\boldsymbol{I}_{N-K} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\begin{bmatrix} \bar{\boldsymbol{U}}_s^T \\ \bar{\boldsymbol{U}}_n^T \\ \bar{\boldsymbol{U}}_o^T \end{bmatrix}, \tag{42}$$

where $\bar{\boldsymbol{\Lambda}}_s = \mathsf{diag}(\bar{\lambda}_1, \cdots, \bar{\lambda}_{K-\tilde{K}})$, with $\bar{\lambda}_k > \sigma^2$; and the columns of $\bar{\boldsymbol{U}}_s$ form an orthonormal basis of the subspace $\mathsf{range}(\boldsymbol{S}) \bigcap \mathsf{null}\left(\tilde{\boldsymbol{S}}^T\right)$. Denote $\tilde{\boldsymbol{d}}_1$ and $\tilde{\boldsymbol{m}}_1$ as respectively the linear decorrelating detector and the linear MMSE detector for the desired user, assuming that only the first $\tilde{K}$ users are present, i.e., [cf.[2]]

$$\tilde{\boldsymbol{d}}_1 \;\triangleq\; \tilde{\boldsymbol{S}}\left(\tilde{\boldsymbol{S}}^T\tilde{\boldsymbol{S}}\right)^{-1}\tilde{\boldsymbol{e}}_1, \tag{43}$$

$$\tilde{\boldsymbol{m}}_1 \;\triangleq\; \tilde{\boldsymbol{S}}\left(\tilde{\boldsymbol{S}}^T\tilde{\boldsymbol{S}} + \sigma^2\boldsymbol{I}_{\tilde{K}}\right)^{-1}\tilde{\boldsymbol{e}}_1. \tag{44}$$

It is shown in [3] that the group-blind linear hybrid detector defined by (26) can also be expressed in terms of the signal subspace components of $\left(\bar{\boldsymbol{P}}\boldsymbol{C}_r\bar{\boldsymbol{P}}\right)$, as

$$\boldsymbol{w}_1 \;=\; \left(\boldsymbol{I}_N - \bar{\boldsymbol{U}}_s\bar{\boldsymbol{\Lambda}}_s^{-1}\bar{\boldsymbol{U}}_s^T\boldsymbol{C}_r\right)\tilde{\boldsymbol{d}}_1, \tag{45}$$

Moreover, it is also shown in [3] that the group-blind linear MMSE detector (Note that by definition this detector is different from the true linear MMSE detector, cf. [3].) is given by

$$\boldsymbol{w}_1 \;=\; \left(\boldsymbol{I}_N - \bar{\boldsymbol{U}}_s\bar{\boldsymbol{\Lambda}}_s^{-1}\bar{\boldsymbol{U}}_s^T\boldsymbol{C}_r\right)\tilde{\boldsymbol{m}}_1. \tag{46}$$

The form-I group-blind detectors are formed by replacing the eigencomponents of $\left(\bar{\boldsymbol{P}}\boldsymbol{C}_r\bar{\boldsymbol{P}}\right)$ in (45) and (46) by those of the corresponding sample estimate $\left(\bar{\boldsymbol{P}}\hat{\boldsymbol{C}}_r\bar{\boldsymbol{P}}\right)$. Denote $\bar{\boldsymbol{S}} \triangleq [\boldsymbol{s}_{\tilde{K}+1}, \cdots, \boldsymbol{s}_K]$. The following result gives the asymptotic distribution of the estimated form-I group-blind linear detectors.

Theorem 3 *Let $\hat{\boldsymbol{C}}_r$ given by (14) be the sample autocorrelation matrix of the received signals based on M samples. Let $\hat{\bar{\boldsymbol{\Lambda}}}_s$ and $\hat{\bar{\boldsymbol{U}}}_s$ contain respectively the largest $(K - \tilde{K})$ eigenvalues of $\left(\bar{\boldsymbol{P}}\hat{\boldsymbol{C}}_r\bar{\boldsymbol{P}}\right)$ and the corresponding eigenvectors. Let $\hat{\boldsymbol{w}}_1$ be the estimated form-I group-blind linear detector, given by*

$$\hat{\boldsymbol{w}}_1 \;=\; \left(\boldsymbol{I}_N - \hat{\bar{\boldsymbol{U}}}_s\hat{\bar{\boldsymbol{\Lambda}}}_s^{-1}\hat{\bar{\boldsymbol{U}}}_s^T\hat{\boldsymbol{C}}_r\right)\boldsymbol{v}, \tag{47}$$

where $\boldsymbol{v} \triangleq \tilde{\boldsymbol{d}}_1$ for the group-blind linear hybrid detector; and $\boldsymbol{v} \triangleq \tilde{\boldsymbol{m}}_1$ for the group-blind linear MMSE detector. Then

$$\sqrt{M}\left(\hat{\boldsymbol{w}}_1 - \boldsymbol{w}_1\right) \;\to\; \mathcal{N}(\boldsymbol{0}, M\cdot\boldsymbol{C}_w), \quad \textit{in distribution, as } M \to \infty,$$

with

$$\boldsymbol{C}_w = \frac{1}{M}\left\{\left(\boldsymbol{w}_1^T\boldsymbol{C}_r\boldsymbol{v}\right)\bar{\boldsymbol{U}}_s\bar{\boldsymbol{\Lambda}}_s^{-1}\bar{\boldsymbol{U}}_s^T - 2\bar{\boldsymbol{U}}_s\bar{\boldsymbol{\Lambda}}_s^{-1}\bar{\boldsymbol{U}}_s^T\bar{\boldsymbol{S}}\bar{\boldsymbol{D}}^2\bar{\boldsymbol{S}}\bar{\boldsymbol{U}}_s\bar{\boldsymbol{\Lambda}}_s^{-1}\bar{\boldsymbol{U}}_s^T + \tau\bar{\boldsymbol{U}}_n\bar{\boldsymbol{U}}_n^T\right\}, \tag{48}$$

where

$$\tau \;=\; \sigma^2\left(\boldsymbol{C}_r\boldsymbol{v}\right)^T\bar{\boldsymbol{U}}_s\bar{\boldsymbol{\Lambda}}_s^{-1}\left(\bar{\boldsymbol{\Lambda}}_s - \sigma^2\boldsymbol{I}_{K-\tilde{K}}\right)^{-2}\bar{\boldsymbol{U}}_s^T\left(\boldsymbol{C}_r\boldsymbol{v}\right), \tag{49}$$

$$\bar{\boldsymbol{D}} \;=\; \mathsf{diag}\left\{A_{\tilde{K}+1}^2\left(\boldsymbol{w}_1^T\boldsymbol{s}_{\tilde{K}+1}\right), \cdots, A_K^2\left(\boldsymbol{w}_1^T\boldsymbol{s}_K\right)\right\}. \tag{50}$$

As before, the SINR's for the form-I group-blind detectors can be expressed in terms of $\boldsymbol{R}$, σ^2 and $\boldsymbol{A}$. However, the closed-form SINR expressions are too complicated for this case and we therefore do not present them here.

5 Simulation Results

In this section, we provide some numerical examples to demonstrate the good match between the analytical performance assessment of various estimated detectors discussed in the previous sections and the simulation results, in terms of both the average output SINR.

We consider a system with $K = 11$ users, of which $\tilde{K} = 7$ are known users. The user spreading sequences are randomly generated with processing gain $N = 13$. All users have the same amplitudes. Figure 1 shows both the analytical and the simulated SINR performance for four different estimated detectors, namely, the DMI blind detector, the subspace blind detector, the form-I group-blind detector and the form-II group-blind detector. For each detector, the SINR is plotted as a function of the number of signal samples (M) used for estimating the detector, at some fixed SNR. It is seen that the agreement between the analytical performance assessment and the simulation results is excellent. The mismatch between the analytical and simulation performance occurs for small values of M, which is not surprising since the analytical performance is based on asymptotic analysis. Moreover, it is seen that the group-blind detectors converges faster than both the subspace blind detector and the DMI blind detector. That is, it requires less samples to reach the asymptotic performance than the two blind detectors.

6 Conclusions

In this paper, we have analyzed the asymptotic performance of several blind and group-blind multiuser detectors. Typically, these detectors are estimated based on a sequence of received signals of length M. Several detectors, including the direct-matrix inversion (DMI) blind detector, the subspace blind detector, the form-I and form-II group-blind hybrid detectors, are considered. We have obtained the asymptotic distributions of these estimated detectors, when the number of signal samples used for detector estimation, M, is large. Based on these asymptotic results, we have obtained expressions of the average output signal-to-interference-plus-noise ratios (SINR's) for these estimated detectors. The analytical results obtained in this paper give insights on the performance difference among these blind and group-blind detectors. Moreover, the analytical performance matches well with the simulation results.

References

[1] M. Honig, U. Madhow, and S. Verdú. Blind adaptive multiuser detection. *IEEE Trans. Inform. Theory*, IT-41(4):944–960, July 1995.

[2] S. Verdú. *Multiuser Detection.* Cambridge University Press: Cambridge, UK, 1998.

[3] X. Wang and A. Høst-Madsen. Group-blind multiuser detection for uplink CDMA. *IEEE J. Select. Areas Commun.*, 17(11), Nov. 1999.

[4] X. Wang and H.V. Poor. Blind equalization and multiuser detection for CDMA communications in dispersive channels. *IEEE Trans. Commun.*, COM-46(1):91–103, Jan. 1998.

[5] X. Wang and H.V. Poor. Blind multiuser detection: A subspace approach. *IEEE Trans. Inform. Theory*, 44(2):677–691, Mar. 1998.

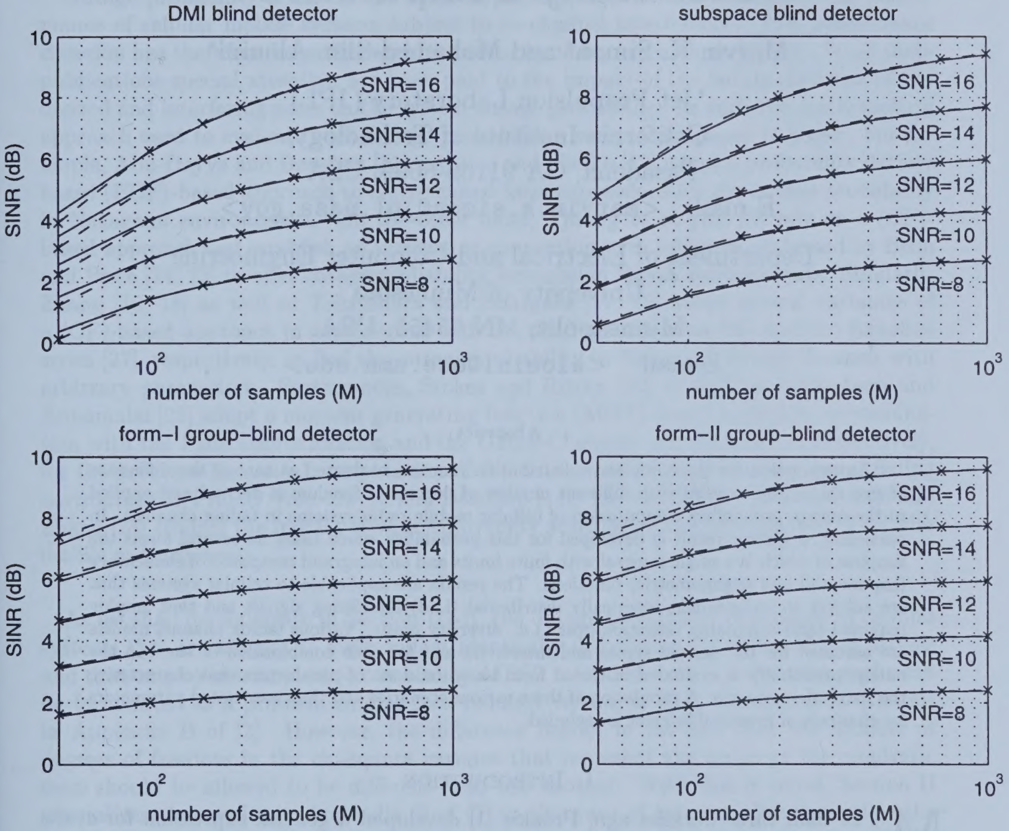


Figure 1: The output average SINR versus the number of signal samples M for different estimated detectors. $N = 13, K = 11, \tilde{K} = 7$. The solid line is the analytical performance, and the dashed line is the simulation performance.

On the Difference of Two Chi-Square Variates with Application to Outage Probability Computation[◇]

Marvin K. Simon[*] and Mohamed-Slim Alouini[⋆]

[*]Jet Propulsion Laboratory (JPL)
California Institute of Technology,
Pasadena, CA 91109-8099, USA.
E-mail: <marvin.k.simon@jpl.nasa.gov>

[⋆]Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, USA.
E-mail: <alouini@ece.umn.edu>

Abstract

An expression for the cumulative distribution function evaluated at zero of the difference of two chi-square variates with different number of degrees of freedom is derived and applied to the outage probability computation of cellular mobile radio systems in fading channels. In particular, a generic result is developed for this probability which takes on several forms the simplest of which is a single integral with finite limits and an integrand composed of elementary (exponential and trigonometric) functions. The results are applicable to cellular systems that are subject to independent identically distributed (i.i.d) interfering signals and that employ maximal-ratio combining reception over i.i.d. diversity paths. Various fading channel models are assumed for the desired signal and interferers and for each combination of the two the outage probability is evaluated in closed form based on a set of parameters that characterize the particular scenario. A tabulation of these various scenarios and their associated parameters for channels of practical interest is included.

I. INTRODUCTION

MORE than three decades ago, Proakis [1] developed a general expression for evaluating the probability of error when the decision variable is a quadratic form in complex-valued Gaussian random variables. Since many problems dealing with the error probability performance of multichannel reception with differentially coherent and noncoherent detection have a decision variable which is in that particular form, the development and results originally obtained in [1] later appeared as Appendix B of [2] and have become a classic in the annals of communication system performance literature. The most general form of the probability of error expression (i.e., [2, Eqn. (B.21)]) obtained by Proakis was given in terms of the first order Marcum Q-function and modified Bessel functions of the first kind. Although implied but not explicitly given in [1, 2], this general form can be rewritten in terms of the generalized Marcum Q-function [3, Eq. (60)] (or equivalently [4, Eq. (9A.11)]), from which one can derive a form corresponding to a single integral with finite limits and an integrand composed of simple functions (exponentials and trigonometrics) [3, Eqs. (62) and (63)] (or equivalently [4, Eq. (9A.20) and Eq. (9A.22)]). It should be pointed out that in applying Proakis' general formula to the error

---

[◇] This work was supported in part by the National Science Foundation (NSF) under Grant CCR-9983462.

probability evaluation of differentially coherent and noncoherent systems ([5, 6, 7, 8, 3] are just a few examples of such recent applications), the case of interest has always been where the number of degrees of freedom in the chi-square variates that represent each term of the quadratic form is the same for all.

Outage probability is one of the most important criterion for evaluating the performance of cellular mobile systems subject to co-channel interference. This performance criterion has therefore been thoroughly studied by many researchers [9]-[23]. In all these publications special attention has been paid to the impact of the fading statistics of the desired and interfering users on the system outage probability. As such the mathematical approach used to evaluate such outage probabilities varies from paper to paper. For example, Abu-Dayya and Beaulieu [11] and Yao and Sheikh [12] adopt a probability density based (PDF)-based approach to derive closed form expressions for the outage probability in Nakagami environments. On the other hand, Tjhung *et al.* [15, 19] rely on a (CF)-based approach and establish an interesting connection with integrals evaluated by Bello and Price [24, 25] to offer outage probability formulas in Rician environments. Similarly, Zhang [16, 18] as well as Tellambura and Bhargava [17, 22] adopt several variantes of a (CF)-based approach in conjunction with the Gil-Palaez lemma [26] and the Beaulieu series [27], respectively, to find the outage probability in Nakagami fading channels with arbitrary parameters. Furthermore, Stokes and Ritcey [20] as well as Tellambura and Annamalai [23] adopt a moment generating function (MGF)-based approach in combination with the Padé approximation and the Gauss-Chebysev approximation, respectively, for the accurate numerical computation of the outage probability in very general fading environments. Finally, Abu-Dayya and Beaulieu [14], Yao and Sheikh [12], and Aalo and Zhang [28] studied the impact of various diversity schemes on the performance of cellular mobile radio systems.

As we shall demonstrate in Section III-B, the outage probability evaluation of cellular systems that are subject to independent identically distributed (i.i.d) interfering signals and that employ maximal-ratio combining (MRC) reception over i.i.d diversity paths can be formulated as a problem similar to a reduced version of that considered by Proakis in Appendix B of [2]. However, the difference resides in the fact that the number of degrees of freedom in the chi-square variates that represent the terms of the quadratic form should be allowed to be different from one another. With this in mind, Section II generalizes the results in Appendix B of [2] to this case of interest. We show that the resulting new expression, which is related to the cumulative distribution function (CDF) of the difference of two chi-square variates with different degrees of freedom, takes on several forms the simplest of which is a single integral with finite limits and an integrand composed of elementary (exponential and trigonometric) functions. Once this result is established, we next apply it in Section III to the outage probability problem arriving at *a generic closed form formula* for its evaluation whereupon we treat specific scenarios of interest corresponding to various combinations of desired users and interferers fading models. When applicable, we compare the results obtained as special cases of our generic result to those previously reported in the literature. Since the number of different fading combinations discussed herein is quite large, numerical results for the outage probability of these combinations and dependence on the various fading parameters are omitted here. Many of the referenced papers cover such numerical illustrations. In addition, our analysis limits its scope to the effect of multipath fading and does not take into account the impact of log-normal shadowing. This can be done by further averaging the presented results over this effect via Gaussian-Hermite quadratures.

II. A Probability Related to the CDF of the Difference of Two Chi-Square Variates with Different Degrees of Freedom

Consider the following quadratic form:

$$D = \sum_{k=1}^{L_1} A\left|X_k\right|^2 + \sum_{k=1}^{L_2} B\left|Y_k\right|^2, \tag{1}$$

where in general the two integers $L_1 > 1$ and $L_2 > 1$ are not necessarily equal, the two real constants A and B are such that $AB < 0$, and the $\{X_k\}_{k=1}^{L_1}$ and $\{Y_k\}_{k=1}^{L_2}$ are all mutually independent complex Gaussian random variables (RVs) with means $\{\overline{X}_k\}_{k=1}^{L_1}$ and $\{\overline{Y}_k\}_{k=1}^{L_2}$, respectively, and with variances

$$\begin{aligned} \mu_{XX} &= \frac{1}{2}E\left\{\left|X_k - \bar{X}_k\right|^2\right\} \\ \mu_{YY} &= \frac{1}{2}E\left\{\left|Y_k - \bar{Y}_k\right|^2\right\}, \end{aligned} \tag{2}$$

which are independent of k. We are interested in the probability P given by

$$P = \Pr\{D < 0\} = \int_{-\infty}^{0} p(D)\,dD, \tag{3}$$

which is the CDF of the variate D evaluated at zero. This problem is analogous to a reduced version of that treated by Proakis in [1], (or equivalently [2, Appendix B]), the exception being that we now allow the two chi-square variates (sums of magnitude square complex Gaussian RVs) in (2) to have different degrees of freedom (L_1 and L_2). In Appendix A of [29] [1], we prove the following theorem.

Theorem 1: The probability P can be expressed in closed form as

$$\begin{aligned} P = {} & Q_1(a,b) - \exp\left[-\tfrac{1}{2}\left(a^2+b^2\right)\right] I_0(ab) \\ & + \frac{I_0(ab)\exp\left[-\frac{1}{2}\left(a^2+b^2\right)\right]}{[1+\eta]^{L_1+L_2-1}} \sum_{k=0}^{L_1-1} \binom{L_1+L_2-1}{k} \eta^k \\ & + \frac{\exp\left[-\frac{1}{2}\left(a^2+b^2\right)\right]}{[1+\eta]^{L_1+L_2-1}} \left[\sum_{n=1}^{L_1-1} I_n(ab) \sum_{k=0}^{L_1-1-n} \binom{L_1+L_2-1}{k} \eta^k \left(\tfrac{b}{a}\right)^n \right. \\ & \left. - \sum_{n=1}^{L_2-1} I_n(ab) \sum_{k=0}^{L_2-1-n} \binom{L_1+L_2-1}{k} \eta^{L_1+L_2-1-k} \left(\tfrac{a}{b}\right)^n \right], \end{aligned} \tag{4}$$

where $Q_1(\cdot,\cdot)$ is the first-order Marcum Q-function [30], $I_n(\cdot)$ is the nth-order modified Bessel function [31, Sect. 8.43], $\binom{L}{k} = \frac{L!}{k!(L-k)!}$ and the parameters a, b, and η are given by

$$a = \sqrt{\frac{\sum_{k=1}^{L_1} A\left|\bar{X}_k\right|^2}{A\mu_{XX} - B\mu_{YY}}} \tag{5}$$

$$b = \sqrt{\frac{-\sum_{k=1}^{L_2} B\left|\bar{Y}_k\right|^2}{A\mu_{XX} - B\mu_{YY}}} \tag{6}$$

---

[1] Due to space limitation, this appendix is omitted here but is included in an extended version of this paper [29].

$$\eta = -\frac{B\,\mu_{YY}}{A\,\mu_{XX}}. \tag{7}$$

Proof: See [29, Appendix A]. ■

Clearly (4) reduces to the Proakis result [2, Eq. (B-21)] when $L_1 = L_2 = L$.

The result in (4) can alternatively be expressed in terms of the generalized Marcum Q-function as was done in [3] and [4, Appendix 9A] for the original Proakis result (i.e., $L_1 = L_2 = L$). In particular, we obtain after much manipulation the following corollary.

Corollary 1: The probability P can be written in terms of the generalized Marcum Q-function as

$$\begin{aligned} P = Q_1(a,b) &- \left[1 - \frac{\sum_{k=0}^{L_1-1}\binom{L_1+L_2-1}{k}\eta^k}{(1+\eta)^{L_1+L_2-1}}\right]\exp\left(-\tfrac{a^2+b^2}{2}\right) I_0(ab) \\ &+ \tfrac{1}{(1+\eta)^{L_1+L_2-1}}\left[\sum_{l=2}^{L_1}\binom{L_1+L_2-1}{L_1-l}\eta^{L_1-l}\left[Q_l(a,b) - Q_1(a,b)\right]\right. \\ &\left. - \sum_{l=2}^{L_2}\binom{L_1+L_2-1}{L_2-l}\eta^{L_1+l-1}\left[Q_l(b,a) - Q_1(b,a)\right]\right] \end{aligned}, \tag{8}$$

where $Q_l(\cdot,\cdot)$ is the lth-order (generalized) Marcum Q-function [32, 33]. As a check note that (8) reverts to [3, Eq. (60)] (or equivalently [4, Appendix 9A, Eq. (A.7)]) for the case $L_1 = L_2 = L$.

As was done previously in [3] and [4, Appendix 9A], the result in (8) can be put in the form of a single finite-range integral using the alternative representation of the generalized Marcum Q-function [32, 34]. In particular, we get after considerable manipulations this second corollary.

Corollary 2: The probability P can be written in the form of a single integral with finite limits and an integrand composed of simple (trigonometric and exponential) functions as

$$P = \frac{\eta^{L_1}}{2\pi(1+\eta)^{L_1+L_2-1}}\int_{-\pi}^{\pi}\left[\frac{f(L_1,L_2;\zeta,\eta;\theta)}{1+2\zeta\sin\theta+\zeta^2}\right]\exp\left(-\frac{b^2}{2}\left[1+2\zeta\sin\theta+\zeta^2\right]\right)d\theta, \quad 0^+ \le \zeta \triangleq a/b < 1, \tag{9}$$

where

$$f(L_1,L_2;\zeta,\eta;\theta) = f_0(L_1,L_2;\zeta,\eta;\theta) + f_1(L_1,L_2;\zeta,\eta;\theta), \tag{10}$$

and[2]

$$\begin{aligned} f_0(L_1,L_2;\zeta,\eta;\theta) &= \left[-\tfrac{(1+\eta)^{L_1+L_2-1}}{\eta^{L_1}} + \sum_{l=1}^{L_1}\binom{L_1+L_2-1}{L_1-l}\eta^{-l} + \sum_{l=1}^{L_2}\binom{L_1+L_2-1}{L_2-l}\eta^{l-1}\right]\zeta(\zeta+\sin\theta) \\ f_1(L_1,L_2;\zeta,\eta;\theta) &= \sum_{l=1}^{L_1}\binom{L_1+L_2-1}{L_1-l}\eta^{-l}\zeta^{-l+1}\left[\cos\left((l-1)\left(\theta+\tfrac{\pi}{2}\right)\right) - \zeta\cos\left(l\left(\theta+\tfrac{\pi}{2}\right)\right)\right] \\ &+ \sum_{l=1}^{L_2}\binom{L_1+L_2-1}{L_2-l}\eta^{l-1}\zeta^{l}\left[\cos\left(l\left(\theta+\tfrac{\pi}{2}\right)\right) - \zeta\cos\left((l-1)\left(\theta+\tfrac{\pi}{2}\right)\right)\right] \end{aligned} \tag{11}$$

As a check, note that (9) with (11) agrees with [3, Eq. (62) and Eq. (63)] (or equivalently [4, Appendix 9A, Eq. (9A.20) and Eq. (9A.22)]) when $L_1 = L_2 = L$.

---

[2] We note that the term $f_0(L_1,L_2;\zeta,\eta;\theta)$ can be shown to be identically equal to zero. We include it here since in [3] the authors had not recognized this fact (which was pointed out to them by L. -F. Tsaur [35]) and thus to allow comparison with the results presented there for $L_1 = L_2 = L$ we choose to keep it here.

III. Application to the Outage Probability Evaluation

A. Fading and System Models

A.1 Channel Fading Models

We consider Rayleigh, Rice, and Nakagami types of fading. Let S denote the instantaneous faded signal power and let $\Omega = E_S[S]$ denote its short-term average. A key observation is that for the three fading models under consideration, S can be viewed as a chi-square variate since it can be written as

$$S = \sum_{n=1}^{N} |X_n|^2, \tag{12}$$

where the $\{X_n\}_{n=1}^N$ are independent complex Gaussian RVs with common mean $E[X_n] = \overline{X}$ and common variance $\mu_{XX} = \frac{1}{2}E\{|X_n - \overline{X}_n|^2\}$. In particular, for Rayleigh fading, S is a central chi-square variate with two degrees of freedom ($N = 1$), $\overline{X} = 0$, and $\mu_{XX} = \Omega/2$. For the Rician case, S is a non-central chi-square variate with two degrees of freedom ($N = 1$), $\overline{X} = \sqrt{\frac{K\,\Omega}{K+1}}$, and $\mu_{XX} = \frac{\Omega}{2(K+1)}$, with $K \geq 0$ denoting the Rician factor. Finally, for the Nakagami case, S can viewed as a central chi-square variate ($N = m$), $\overline{X} = 0$, and $\mu_{XX} = \frac{\Omega}{2m}$, where m is the fading parameter which in general can be any real number greater or equal to $1/2$ but which under this model is restricted to integer values.

A.2 Desired and Interference Signals Model

We consider a cellular mobile radio system in which the desired and interfering signals may have different fading statistics. The desired signal is assumed to be received over L independent identically distributed (i.i.d) diversity (time, frequency, or space) paths and to be independent of the interfering signals. The corresponding L received desired instantaneous signal powers $\{S_{D,l}\}_{l=1}^L$ are assumed to have the same average fading power Ω_D and subject to slowly-varying flat Rayleigh, Rician (with Rician factor K_D), or Nakagami (with fading parameter m_D) type of fading. The L replicas of the desired signal are combined as per the rule of MRC and the resulting total desired signal power S_D is thus given by

$$S_D = \sum_{l=1}^{L} S_{D,l} \tag{13}$$

On the other hand, the N_I active interfering signal powers $\{S_{I,n}\}_{n=1}^{N_I}$ are assumed to be i.i.d. with the same short-term average fading power Ω_I and subject to slowly-varying flat Rayleigh, Rician (with Rician factor K_I), or Nakagami (with fading parameter m_I) type of fading. We assume throughout our study that the co-channel interfering signals add up incoherently since this leads to a more realistic assessment of the co-channel interference in cellular systems [36]. Under this assumption the total instantaneous interference power S_I is just the sum of the instantaneous powers of the N_I active interferers and is therefore given by

$$S_I = \sum_{n=1}^{N_I} S_{I,n}. \tag{14}$$

B. A Generic Formula for the Outage Probability

We consider interference limited systems for which an outage is declared when the carrier-to-interference power ratio (CIR) $\lambda = S_D/S_I$ falls below a predetermined protection ratio λ_{th}. Hence, the outage probability P_{out} is given by

$$P_{\text{out}} = P\left[\lambda = \frac{S_D}{S_I} \leq \lambda_{\text{th}}\right], \tag{15}$$

or equivalently

$$P_{\text{out}} = P\left[D = S_D - \lambda_{\text{th}}\ S_I \leq 0\right]. \tag{16}$$

In view of the key observation of Section III-A.1 and under the diversity paths and interference signals i.i.d. assumptions discussed in Section III-A.2, the $\{S_{D,l}\}_{l=1}^{L}$ are chi-square variates and so are the $\{S_{I,n}\}_{n=1}^{N_I}$. Since the sum of i.i.d. chi-square variates is also a chi-square variate, we can conclude from (13) and (14) that S_D and S_I are chi-square variates, respectively. Hence, the RV D as given in (16) can be viewed as the difference of two chi-square variates (with not necessarily the same number of degrees of freedom) and the outage probability, as given in (16), is therefore in the form of the probability P in (3). To obtain this outage probability in closed form, we just need to set $A = 1$ and $B = -\lambda_{\text{th}}$ in (1), whereas the other relevant parameters in (1) (i.e., L_1, L_2, $\bar{X}_k$, $\bar{Y}_k$, μ_{XX}, and μ_{YY}) are set to different values according to the scenario of interest. The four scenarios under consideration in this paper, namely (i) both desired and interfering signals are subject to Nakagami fading (Nakagami/Nakagami), (ii) both desired and interfering signals are subject to Rician fading (Rice/Rice), (iii) desired signal subject to Rician fading whereas interfering signal subject to Nakagami fading (Rice/Nakagami), and (iv) desired signal subject to Nakagami fading whereas interfering signal subject to Rician fading (Nakagami/Rice), are discussed in detail in what follows.

B.1 Nakagami/Nakagami Scenario

To find the outage probability in the case where both the desired and interfering signals are subject to Nakagami fading, we just need to set $L_1 = m_I N_I$, $L_2 = m_d L$, $\bar{X}_k = \bar{Y}_k = 0$, $\mu_{XX} = \frac{\Omega_I}{2m_I}$, and $\mu_{YY} = \frac{\Omega_D}{2m_D}$. Substituting these settings along with $A = 1$ and $B = -\lambda_{\text{th}}$ in (5), (6), and (7), leads to

$$\begin{aligned} a &= 0 \\ b &= 0 \\ \eta &= \frac{m_I \Omega_d}{m_d \lambda_{\text{th}} \Omega_I}. \end{aligned} \tag{17}$$

Substituting (17), $L_1 = m_I N_I$, and $L_2 = m_d L$ in (8) (or equivalently (4) or (9)) and using the fact that $Q_l(0,0) = 1$, $(l = 1, 2, \cdots L)$ the outage probability for the Nakagami/Nakagami scenario reduces to

$$P_{\text{out}} = \frac{1}{\left(1 + \frac{\Omega_d m_I}{\lambda_{th} \Omega_I m_d}\right)^{m_d L + m_I N_I - 1}} \sum_{k=0}^{m_I N_I - 1} \binom{m_d L + m_I N_I - 1}{k} \left(\frac{\Omega_d m_I}{\lambda_{\text{th}} \Omega_I m_d}\right)^k. \tag{18}$$

For the no diversity case (i.e., $L = 1$), (18) is equivalent to the Yao and Sheikh result [12, Eq. (13)] (as we show in Appendix B) and is the same form as the Abu-Dayya and Beaulieu formula [11, Eq. (7a)] for the further particular case of $m_D = m_I = m$. Furthermore, when the desired user is subject to Rayleigh fading ($m_D = 1$), it can be shown using the binomial expansion, that (18) reduces to the compact Tellambura and Bhargava expression [17, Eq. (18)]

$$P_{\text{out}} = 1 - \left(1 + \frac{\lambda_{\text{th}} \Omega_I}{m_I \Omega_d}\right)^{-m_I N_I}, \tag{19}$$

which further simplifies to the Sowerby and Williamson formula given by [9, Eq. (20)]

$$P_{\text{out}} = 1 - \left(1 + \frac{\lambda_{\text{th}}\Omega_I}{\Omega_d}\right)^{-N_I}, \tag{20}$$

when the interferers are also subject to Rayleigh fading (i.e., $m_I = 1$).

B.2 Rice/Rice Scenario

To find the outage probability in the case where both the desired and interfering signals are subject to Rice fading, we just need to set $L_1 = N_I$, $L_2 = L$, $\bar{X}_k = \sqrt{\frac{K_I\,\Omega_I}{K_I+1}}$, $\bar{Y}_k = \sqrt{\frac{K_D\,\Omega_D}{K_D+1}}$, $\mu_{XX} = \frac{\Omega_I}{2(K_I+1)}$ and $\mu_{YY} = \frac{\Omega_D}{2(K_D+1)}$. Substituting these settings along with $A = 1$ and $B = -\lambda_{\text{th}}$ in (5), (6), and (7), leads to

$$\begin{aligned} a &= \sqrt{\frac{2N_I(K_d+1)K_I}{K_d+1+(K_I+1)\frac{\Omega_d}{\Omega_I\lambda_{\text{th}}}}} \\ b &= \sqrt{\frac{2LK_d(K_I+1)}{K_I+1+(K_d+1)\frac{\lambda_{\text{th}}\Omega_I}{\Omega_d}}} \\ \eta &= \frac{(K_I+1)\Omega_d}{(K_d+1)\lambda_{\text{th}}\Omega_I}. \end{aligned} \tag{21}$$

Substituting (21), $L_1 = N_I$, $L_2 = L$ in (4) (or equivalently (4) or (9)) yields the outage probability for the Rice/Rice scenario. For the no diversity case (i.e., $L = 1$), it can be easily shown that the resulting expression is in agreement with the Tjhung *et al.* result [19, Eq. (13)] [3], and it therefore reduces to all the special cases identified and discussed in [19, Section II-B] (in particular the Rice/Rayleigh case studied by Yao and Sheikh in [10] and the Rayleigh/Rice case studied by Haug and Ucci in [13]).

B.3 Rice/Nakagami Scenario

To find the outage probability in the case where the desired user is subject to Rician fading whereas the interfering signals are subject to Nakagami fading, we just need to set $L_1 = m_I N_I$, $L_2 = L$, $\bar{X}_k = 0$, $\bar{Y}_k = \sqrt{\frac{K_D\,\Omega_D}{K_D+1}}$, $\mu_{XX} = \frac{\Omega_I}{2m_I}$ and $\mu_{YY} = \frac{\Omega_D}{2(K_D+1)}$. Substituting these settings along with $A = 1$ and $B = -\lambda_{\text{th}}$ in (5), (6), and (7), leads to

$$\begin{aligned} a &= 0 \\ b &= \sqrt{\frac{2LK_d}{1+\frac{\lambda_{\text{th}}\Omega_I(K_d+1)}{m_I\Omega_d}}} \\ \eta &= \frac{m_I\Omega_d}{\lambda_{th}\Omega_I(K_d+1)}. \end{aligned} \tag{22}$$

Substituting (22), $L_1 = m_I N_I$, and $L_2 = L$ in (8) (or equivalently (4) or (9)) and using the classical identities of the generalized Marcum Q-function [33, Eq. (9)]

$$\begin{aligned} Q_l(0,w) &= e^{-w^2/2}\sum_{k=0}^{l-1}\frac{(w^2/2)^k}{k!}, & l &= 1,2,\cdots,L. \\ Q_l(u,0) &= 1, & l &= 1,2,\cdots,L. \end{aligned} \tag{23}$$

---

[3] More precisely the resulting expression is in agreement with the Tjhung *et al.* result [19, Eq. (13)] if a minor typo is corrected in [19, Eqn. (13)] (the square root of the Bessel function argument should not extend over the denominator of the fraction).

the outage probability for the Rice/Nakagami scenario simplifies to

$$\begin{aligned} P_{\text{out}} &= \frac{e^{-b^2/2}}{(1+\eta)^{m_I N_I+L-1}} \left[\sum_{k=0}^{m_I N_I-1} \binom{m_I N_I + L - 1}{k} \eta^k \right. \\ &+ \left. \sum_{k=0}^{m_I N_I-2} \binom{m_I N_I + L - 1}{k} \eta^k \sum_{n=1}^{m_I N_I-k-1} \frac{(b^2/2)^n}{n!} \right], \end{aligned} \tag{24}$$

which can be shown to be in agreement with the Lin *et al.* expression [21, Eq. (14)] for the no diversity case (i.e., $L = 1$) and with the Yao and Sheikh result [10, Eq. (13)] for the no diversity case (i.e., $L = 1$) and Rayleigh interferers (i.e., $m_I = 1$).

B.4 Nakagami/Rice Scenario

To find the outage probability in the case where the desired user is subject to Nakagami fading whereas the interfering signals are subject to Rician fading, we just need to set $L_1 = N_I$, $L_2 = m_D L$, $\bar{X}_k = \sqrt{\frac{K_I\,\Omega_I}{K_I+1}}$, $\bar{Y}_k = 0$, $\mu_{XX} = \frac{\Omega_I}{2(K_I+1)}$ and $\mu_{YY} = \frac{\Omega_D}{2m_D}$. Substituting these settings along with $A = 1$ and $B = -\lambda_{\text{th}}$ in (5), (6), and (7), leads to

$$\begin{aligned} a &= \sqrt{\frac{2K_I N_I}{1 + (1+K_I)\frac{\Omega_d}{m_d \lambda_{\text{th}} \Omega_I}}} \\ b &= 0 \\ \eta &= \frac{(K_I+1)\Omega_d}{m_d \lambda_{\text{th}} \Omega_I}. \end{aligned} \tag{25}$$

Substituting (25), $L_1 = N_I$, and $L_2 = m_D L$ in (8) (or equivalently (4) or (9)) and using the classical identities of the generalized Marcum Q-function given in (23) leads to the outage probability for the Nakagami/Rice scenario in the following simplified form

$$\begin{aligned} P_{\text{out}} &= 1 - \frac{e^{-a^2/2}}{(1+\eta)^{m_d L+N_I-1}} \left[(1+\eta)^{m_d L+N_I-1} - \sum_{k=0}^{N_I-1} \binom{m_d L + N_I - 1}{k} \eta^k \right. \\ &+ \left. \sum_{k=0}^{m_d L-2} \binom{m_d L + N_I - 1}{k} \eta^{m_d L+N_I-k-1} \sum_{n=1}^{m_d L-k-1} \frac{(a^2/2)^n}{n!} \right]. \end{aligned} \tag{26}$$

IV. CONCLUSION

In this paper we adapted the proof of the classical Proakis result on the probability of error when the decision variable is a quadratic form in complex-valued Gaussian RVs to obtain an expression (given by (4)) for the CDF evaluated at zero of the difference of two chi-square variates with different number of degrees of freedom. We showed that this expression takes on several forms the simplest of which (given by (9)) is a single integral with finite limits and an integrand composed of elementary (exponential and trigonometric) functions. We then exploited this result to obtain a generic expression for the outage probability of cellular mobile radio systems in fading channels. This expression is applicable to cellular systems that are subject to i.i.d. interfering signals and that employ MRC reception over i.i.d. diversity paths. We considered various fading model combinations for the desired signal and interferers and for each combination of the two we evaluated the outage probability in closed form based on a set of parameters that characterize the particular scenario. A tabulation of these various scenarios and their associated parameters for channels of practical interest is summarized in Table I and should be used in conjunction with (8) (or equivalently (4) or (9)) to obtain the outage probability for the various combinations under consideration.

TABLE I

PARAMETERS L_1, L_2, a, b, AND η FOR THE VARIOUS SCENARIOS UNDER CONSIDERATION

Desired/Interferers	L_1	L_2	a	b	η
Rayleigh/Rayleigh	N_I	L	0	0	$\frac{\Omega_d}{\lambda_{\text{th}}\Omega_I}$
Rayleigh/Rice	N_I	L	$\sqrt{\frac{2K_I N_I}{1+(1+K_I)\frac{\Omega_d}{\lambda_{\text{th}}\Omega_I}}}$	0	$\frac{(K_I+1)\Omega_d}{\lambda_{\text{th}}\Omega_I}$
Rayleigh/Nakagami	$m_I N_I$	L	0	0	$\frac{m_I\Omega_d}{\lambda_{\text{th}}\Omega_I}$
Rice/Rayleigh	N_I	L	0	$\sqrt{\frac{2LK_d}{1+(K_d+1)\frac{\lambda_{\text{th}}\Omega_I}{\Omega_d}}}$	$\frac{\Omega_d}{(K_d+1)\lambda_{\text{th}}\Omega_I}$
Rice/Rice	N_I	L	$\sqrt{\frac{2N_I(K_d+1)K_I}{K_d+1+(K_I+1)\frac{\Omega_d}{\Omega_I\lambda_{\text{th}}}}}$	$\sqrt{\frac{2LK_d(K_I+1)}{K_I+1+(K_d+1)\frac{\lambda_{\text{th}}\Omega_I}{\Omega_d}}}$	$\frac{(K_I+1)\Omega_d}{(K_d+1)\lambda_{\text{th}}\Omega_I}$
Rice/Nakagami	$m_I N_I$	L	0	$\sqrt{\frac{2LK_d}{1+\frac{(1+K_d)}{m_I}\lambda_{\text{th}}\Omega_I\Omega_d}}$	$\frac{m_I\Omega_d}{(1+K_d)\lambda_{\text{th}}\Omega_I}$
Nakagami/Rayleigh	N_I	$m_d L$	0	0	$\frac{\Omega_d}{m_d\lambda_{\text{th}}\Omega_I}$
Nakagami/Rice	N_I	$m_d L$	$\sqrt{\frac{2K_I N_I}{1+(1+K_I)\frac{\Omega_d}{m_d\lambda_{\text{th}}\Omega_I}}}$	0	$\frac{(K_I+1)\Omega_d}{m_d\lambda_{\text{th}}\Omega_I}$
Nakagami/Nakagami	$m_I N_I$	$m_d L$	0	0	$\frac{m_I\Omega_d}{m_d\lambda_{\text{th}}\Omega_I}$

REFERENCES

[1] J. G. Proakis, "On the probability of error for multichannel reception of binary signals," *IEEE Trans. Commun. Technol.*, vol. COM-16, pp. 68–71, Febuary 1968.

[2] J. G. Proakis, *Digital Communications*. New York, NY: McGraw-Hill, third ed., 1995.

[3] M. K. Simon and M. -S. Alouini, "A unified approach to the probability of error for noncoherent and differentially coherent modulations over generalized fading channels," *IEEE Trans. Commun.*, vol. COM-46, pp. 1625–1638, December 1998.

[4] M. K. Simon and M. -S. Alouini, "Digital communications over generalized fading channels: A unified approach to performance analysis." New York, NY: John Wiley & Sons, to be published in July 2000.

[5] Y. C. Chow, J. P. McGeehan, and A. R. Nix, "Simplified error bound analysis for M-DPSK in fading channels with diversity reception," *IEE Proc.-Commun.*, vol. 141, pp. 341–350, October 1994.

[6] D. L. Noneaker and M. B. Pursley, "Error probability bounds for M-PSK and M-DPSK and selective fading diversity channels," *IEEE Trans. Veh. Technol.*, vol. VT-43, pp. 997–1005, November 1994.

[7] X. Dong, T. T. Tjhung, and F. Adachi, "Error probability analysis for 16 STAR-QAM in frequency-selective Rician fading with diversity reception," *IEEE Trans. Veh. Technol.*, vol. VT-47, pp. 924–935, August 1998.

[8] I. Korn, "Binary PRCPM with differential phase detection and maximal ratio combining diversity in fading channels," *IEEE Trans. Veh. Technol.*, vol. VT-47, pp. 936–946, August 1998.

[9] K. W. Sowerby and A. G. Williamson, "Outage probability calculations for multiple cochannel interferers in cellular mobile radio systems," *IEE Proc. (Pt. F)*, vol. 135, pp. 208–215, June 1988.

[10] Y. D. Yao and A. U. H. Sheikh, "Outage probability analysis for microcell mobile radio systems with co-channel interferers in Rician/Rayleigh fading environment," *IEE Electron. Lett.*, vol. 26, pp. 864–866, June 1990.

[11] A. A. Abu-Dayya and N. C. Beaulieu, "Outage probabilities of cellular mobile radio systems with multiple Nakagami interferers," *IEEE Trans. Veh. Technol.*, vol. VT-40, pp. 757–768, November 1991.

[12] Y. -D. Yao and A. U. H. Sheikh, "Investigation into cochannel interference in microcellular mobile radio systems," *IEEE Trans. Veh. Technol.*, vol. VT-41, pp. 114–123, May 1992.

[13] J. R. Haug and D. R. Ucci, "Outage probability for microcellular radio systems in a Rayleigh/Rician fading environment," in *Proc. IEEE Int. Conf. on Commun. (ICC'92)*, pp. 312.4.1–312.4.5, June 1992.

[14] A. A. Abu-Dayya and N. C. Beaulieu, "Outage probabilities of diversity cellular systems with cochannel inteference in Nakagami fading," *IEEE Trans. Veh. Technol.*, vol. VT-41, pp. 343–355, November 1992.

[15] T. T. Tjhung, C. C. Chai, and X. Dong, "Outage probability for a Rician signal in L Rician interferers," *IEE Electron. Lett.*, vol. 31, pp. 532–533, March 1995.

[16] Q. T. Zhang, "Outage probability of cellular mobile radio in the presence of multiple Nakagami interferers with arbitrary fading parameters," *IEEE Trans. Veh. Technol.*, vol. VT-44, pp. 661–667, August 1995.

[17] C. Tellambura and V. K. Bhargava, "Outage probability analysis for cellular mobile radio systems subject to Nakagami fading and shadowing," *Trans. IECE of Japan*, vol. E 78-B, pp. 1416–1423, October 1995.

[18] Q. T. Zhang, "Outage probability in cellular mobile radio due to Nakagami signal and interferers with arbitrary parameters," *IEEE Trans. on Veh. Technol.*, vol. 45, pp. 364–372, May 1996.

[19] T. T. Tjhung, C. C. Chai, and X. Dong, "Outage probability for lognormal-shadowed Rician channels," *IEEE Trans. Veh. Technol.*, vol. VT-46, pp. 400–407, May 1997.

[20] J. W. Stokes and J. A. Ritcey, "A general method for evaluating outage probabilities using Pade approxima-

tions," in *Proc. IEEE Global Commun. Conf. GLOBECOM'98, Sydney, Australia*, pp. 1485–1490, November 1998.

[21] J. -C. Lin, W. -C. Kao, Y. T. Su, and T. -H. Lee, "Outage and coverage considerations for microcellular mobile radio systems in a shadowed-Rician/shadowed-Nakagami environment," *IEEE Trans. Veh. Technol.*, vol. VT-48, pp. 66–75, January 1999.

[22] C. Tellambura, "Cochannel interference computation for arbitrary Nakagami fading," *IEEE Trans. on Veh. Technol.*, vol. 48, pp. 487–489, March 1999.

[23] C. Tellambura and A. Annamalai, "A unified numerical approach for computing the outage probability for mobile radio systems," *IEEE Commun. Letters*, vol. 3, pp. 97–99, April 1999.

[24] P. A. Bello, "Binary error probabilities over selectively fading channels containing specular components," *IEEE Trans. Commun. Technol.*, vol. COM-14, pp. 400–406, August 1966.

[25] R. Price, "Some noncentral F-distributions expressed in closed form," *Biometrika*, vol. 51, pp. 107–122, 1964.

[26] J. Gil-Pelaez, "Notes on the inversion theorem," *Biometrika*, vol. 38, pp. 481–482, 1951.

[27] N. C. Beaulieu, "An infinite series for the computation of the complementary probability distribution function of a sum of independent random variables and its application to the sum of Rayleigh random variables," *IEEE Trans. Commun.*, vol. COM-26, pp. 1463–1474, September 1990.

[28] V. A. Aalo and J. Zhang, "On the effect of cochannel interference on average error rates in Nakagami-fading channels," *IEEE Commun. Letters*, vol. 3, pp. 136–138, May 1999.

[29] M. K. Simon and M. -S. Alouini, "On the difference of two chi-square variate with application to outage probability computation." Submitted to *IEEE Trans. Commun.*, May 2000.

[30] J. I. Marcum, "Table of Q functions," tech. rep., U. S. Air Force Project RAND Research Memorandum M-339, ASTIA Document AD 1165451, Rand Corporation, Santa Monica, CA, January 1950.

[31] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* San Diego, CA: Academic Press, fifth ed., 1994.

[32] C. W. Helstrom, *Elements of Signal Detection and Estimation.* Englewood Cliffs, NJ: PTR Prentice-Hall, 1995.

[33] G. M. Dillard, "Recursive computation of the generalized Q-function," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-9, pp. 614–615, July 1973.

[34] M. K. Simon, "A new twist on the Marcum Q-function and its applications," *IEEE Commun. Letters*, vol. 2, pp. 39–41, February 1998.

[35] L. -F. Tsaur, Private communication.

[36] R. Prasad and A. Kegel, "Improved assessement of interference limits in cellular radio performance," *IEEE Trans. Veh. Technol.*, vol. VT-40, pp. 412–419, May 1991.

Low-Complexity Bandwidth-Efficient MUI/ISI-resilient CDMA based on Block-Spreading*

Shengli Zhou, Georgios B. Giannakis and Christophe Le Martret

Dept. of Elec. and Comp. Engr., Univ. of Minnesota
200 Union Street SE, Minneapolis, MN 55455
Emails: `{szhou,georgios,lemartre}@ece.umn.edu`
Tel: (612) 626-7781, Fax: (612) 625-4583

Abstract

A low-complexity MUI-free CDMA transceiver is developed in this paper. Relying on block spreading and zero padded transmissions, mutual orthogonality between different users' spreading codes is maintained at the receiver even after frequency selective propagation, which leads to deterministic multiuser separation with a low-complexity code-matching front-end. In addition to MUI-free reception, the proposed system guarantees channel-irrespective symbol recovery. Furthermore, it achieves high bandwidth efficiency by increasing the symbol block size. Filling the zero-gaps with known symbols allows for perfectly constant modulus transmissions. Simulation results demonstrate improved performance of the proposed system relative to competing alternatives.

1 Introduction

Relying on orthogonal spreading codes, Code Division Multiple Access (CDMA) systems allow simultaneous transmissions from multiple users over the same bandwidth. However, when the chip rate increases, the underlying multipath channel becomes frequency selective; it introduces inter chip interference (ICI) and thus destroys code orthogonality at the receiver. The latter gives rise to Multiuser Interference (MUI). To suppress MUI, various multiuser detectors are available [7], e.g., the linear decorrelator or Zero Forcing (ZF), the Minimum Mean Square Error (MMSE), as well as the nonlinear Decision Feedback (DF) and Maximum Likelihood (ML) receivers. However, these schemes require knowledge of the multipath channels for all users and/or suppress MUI statistically (except for the ZF option) even with exact Channel State Information (CSI). In addition to increased complexity that comes with multichannel estimation and multiuser detection, there even exist frequency-selective channels preventing symbol detection no matter what receiver is used [8].

To remove MUI deterministically *regardless* of the underlying FIR multipath channels, several alternatives have been proposed recently. Those include the Orthogonal Frequency Division Multiple Access (OFDMA) [5] (and generalizations [6]), where complex exponentials are utilized as information-bearing subcarriers and thus retain their shape and orthogonality when passing through multipath channels. However, when the channels have nulls (or deep fades) on some

---

*This research is supported by the NSF Wireless Initiative grant no. 99-79443 and the ARL grant no. DAAL01-98-Q-0648

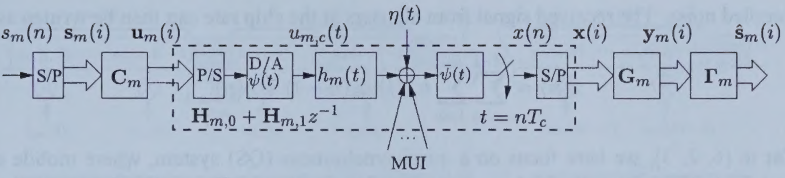


Figure 1: Continuous and discrete-time equivalent system model (mth user shown)

subcarriers, the information symbols on those subcarriers will be lost. Therefore, OFDMA-like transceivers require extra diversity (such as frequency hopping or channel coding) to ameliorate fading effects. To guarantee channel-irrespective MUI-free reception and symbol recovery, A Mutually-Orthogonal Usercode-Receiver (AMOUR) system was proposed in [2]. However, similar to other multicarrier systems, AMOUR transmissions are not Constant-Modulus (C-M) in general, even though a special code design can be C-M which will turn out to fall under the system designs developed herein. AMOUR codes are generally complex valued and bandwidth efficiency drops by 50% if real codes are to be designed from complex ones [8]. To maintain C-M at the transmitter and facilitate low-complexity receivers for MUI-free reception, the so-called shift orthogonal codes (which are not only orthogonal to each other but also to their shifted versions) were proposed in [3, 4]. However, to maintain shift orthogonality, a 50% bandwidth efficiency penalty is paid for both the real and the complex codes in [3, 4].

In this paper, we develop novel MUI-free CDMA transceivers that rely on block-spread constant-modulus transmissions. Thanks to block spreading and zero padding, mutual orthogonality between different users' codes is preserved even after multipath propagation, which allows for deterministic multiuser separation through a low-complexity code-matching front-end (Section 3). In addition to MUI-free reception, channel-irrespective symbol recovery is also guaranteed. Since the only requirement is mutual orthogonality among users' codes, the code design is very flexible and enables fast algorithms at the receiver. By increasing the symbol block size, the proposed system achieves high bandwidth efficiency. Perfectly constant modulus transmissions become available by filling the zero gaps with known symbols. Simulations are then performed in Section 4 to corroborate the improved performance of the proposed system.

2 System Modeling

The block diagram in Fig. 1 describes a CDMA system model in the downlink or uplink scenario, where only one user (the mth user out of a maximum M users) is shown. Unlike traditional spreading which is performed over a *single symbol*, we here use block spreading that operates on a *block of symbols*; in other words, the information stream of the mth user $s_m(n)$ is first parsed into K-long blocks $\mathbf{s}_m(i) := [s(iK), s(iK+1), \ldots, s(iK+K-1)]^T$ and then block spread by a $P \times K$ spreading matrix $\mathbf{C}_m$ to obtain the $P \times 1$ output vector $\mathbf{u}_m(i) := \mathbf{C}_m\mathbf{s}_m(i)$. After Parallel to Serial (P/S) conversion, the mth user's coded chip sequence $u_m(n)$ is pulse shaped to the corresponding continuous time signal $u_m(t)$. The latter propagates through a (possibly *unknown*) channel $h_m(t)$ and is filtered by the receive filter. Let $h_m(l)$ (with order L_m) denote the equivalent discrete time channel impulse response that includes the mth user's asynchronism in the form of delay factors as well as transmit-receive filters and multipath effects, and $\eta(n)$ be the sampled noise. The received signal from all users at the chip rate can then be written as

$$x(n) = \sum_{m=0}^{M-1} \sum_{l=0}^{L_m} h_m(l) u_m(n-l) + \eta(n). \tag{1}$$

Similar to [6, 2, 3], we here focus on a quasi-synchronous (QS) system, where mobile users attempt to synchronize with the base-station's pilot waveform. Thus, the maximum channel order $L := \max_m L_m$ for all users can be made small (relative to the block size K) provided that we select K and the transmitted block length $P \gg L$.

The received samples $x(n)$ are serial to parallel converted to form $P \times 1$ vectors: $\mathbf{x}(i) := [x(iP), x(iP+1), \ldots, x(iP+P-1)]^T$. Define the noise vector $\boldsymbol{\eta}(i) := [\eta(iP), \eta(iP+1), \ldots, \eta(iP+P-1)]^T$; and, let $\mathbf{H}_{m,0}$ be the $P \times P$ lower triangular Toeplitz matrix with first column $[h_m(0), \ldots, h_m(L), 0, \ldots, 0]^T$, and $\mathbf{H}_{m,1}$ be the $P \times P$ upper triangular Toeplitz matrix with first row $[0, \ldots, 0, h_m(L), \ldots, h_m(1)]$. The Input-Output block relationship of the channels can be described by [8]:

$$\mathbf{x}(i) = \sum_{m=0}^{M-1} [\mathbf{H}_{m,0}\mathbf{C}_m \mathbf{s}_m(i) + \mathbf{H}_{m,1}\mathbf{C}_m \mathbf{s}_m(i-1)] + \boldsymbol{\eta}(i), \tag{2}$$

where the second term in the sum accounts for the Inter Block Interference (IBI).

To extract the μ-th user from $\mathbf{x}(i)$, a linear multiuser separating front-end for user μ, denoted by the matrix $\mathbf{G}_\mu$, is applied to $\mathbf{x}(i)$. The MUI-free output is then equalized by a linear single user equalizer $\boldsymbol{\Gamma}_\mu$ as in the following:

$$\mathbf{y}_\mu(i) = \mathbf{G}_\mu \mathbf{x}(i), \qquad \hat{\mathbf{s}}_\mu(i) = \boldsymbol{\Gamma}_\mu \mathbf{y}_\mu(i). \tag{3}$$

In the next section, we will design the transceiver pairs $\{\mathbf{C}_m, \mathbf{G}_m\}_{m=0}^{M-1}$, to separate the superposition of multiuser signals deterministically regardless of multipath propagation of (quasi-) synchronous transmissions through FIR channels of maximum order L.

3 Transceiver design

For each of the M users, we assign a distinct $M \times 1$ vector $\mathbf{c}_m := [c_m(0), \ldots, c_m(M-1)]^T$. These vectors are selected to be mutually orthogonal which can be formally stated as a design constraint ($\delta(\cdot)$ stands for Kronecker's delta):
d1) select code-generating vectors: $\mathbf{c}_\mu^{\mathcal{H}} \mathbf{c}_m = \delta(m-\mu)$, $\forall m, \mu \in [0, M-1]$.

Let $\otimes$ denote the Kronecker's product, $\mathbf{I}_{K+L}$ stand for the $(K+L) \times (K+L)$ identity matrix and define the $(K+L) \times K$ zero-padding (ZP) matrix $\mathbf{T}_{zp} := [\mathbf{I}_K, \mathbf{0}_{K \times L}]^T$. With these notations, we design our $P \times K$ user code matrices $\mathbf{C}_m$ and correspondingly the $(K+L) \times P$ separating matrices $\mathbf{G}_m$ as $((\cdot)^{\mathcal{H}}$ stands for conjugate transpose):

$$\mathbf{C}_m = \mathbf{c}_m \otimes \mathbf{T}_{zp}, \qquad \mathbf{G}_m = \mathbf{c}_m^{\mathcal{H}} \otimes \mathbf{I}_{K+L}, \tag{4}$$

where using the Kronecker product definition each block-spreading code (column of $\mathbf{C}_m$) turns out to have length $P = M(K+L)$. For the $\mathbf{C}_m$ defined in (4), the serial version of the transmitted block $\mathbf{u}_m(i) = \mathbf{C}_m \mathbf{s}_m(i)$ is depicted in Fig. 2 (upper part). The roles of $\mathbf{s}_m(i)$ and $\mathbf{c}_m$ are interchanged when one compares symbol spreading (lower part of Fig. 2) with block spreading

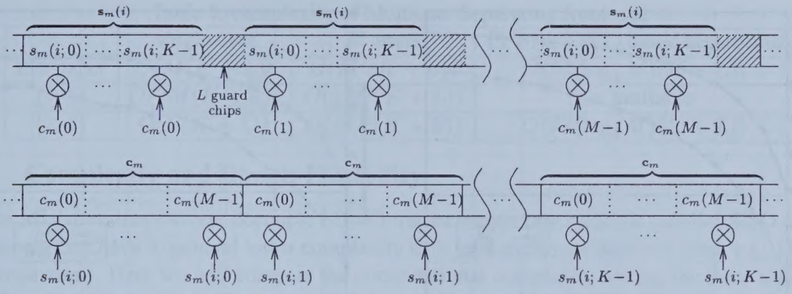


Figure 2: (Upper part) serial version of the ith transmitted block $\mathbf{u}_m(i) = \mathbf{C}_m \mathbf{s}_m(i)$ of length $M(K+L)$ versus the conventional CDMA transmission of K symbols, each spread with a short code of length M (lower part).

(upper part of Fig. 2). Indeed, the code in symbol-periodic spreading spreads each symbol, whereas the same block of symbols spreads each chip of the code in block spreading.

Note that the receive matrix $\mathbf{G}_m$ in (4) depends only on the user's code and constitutes the code-matching user-separating front-end for user m. Next, we will show how this simple user-specific matrix $\mathbf{G}_m$ eliminates MUI deterministically (by design) regardless of *unknown* FIR frequency-selective multipath channels.

Let $\bar{\mathbf{H}}_{m,0}$ be the $(K+L) \times (K+L)$ lower triangular Toeplitz matrix with first column $[h_m(0), \ldots, h_m(L), 0, \ldots, 0]^T$, $\bar{\mathbf{H}}_{m,1}$ the $(K+L) \times (K+L)$ upper triangular Toeplitz matrix with first row $[0, \ldots, 0, h_m(L), \ldots, h_m(1)]$, and $\mathbf{J}_M$ an $M \times M$ shift matrix which is defined as the lower triangular Toeplitz matrix with first column $[0, 1, 0, \ldots, 0]^T$. We can then split the $P \times P$ matrix $\mathbf{H}_{m,0}$ into smaller blocks and rewrite it as:

$$\mathbf{H}_{m,0} = \mathbf{I}_M \otimes \bar{\mathbf{H}}_{m,0} + \mathbf{J}_M \otimes \bar{\mathbf{H}}_{m,1}. \tag{5}$$

At this point, let us recall the identity of Kronecker products applied to matrices (with matching dimensions): $(\mathbf{A}_1 \otimes \mathbf{A}_2)(\mathbf{A}_3 \otimes \mathbf{A}_4) = (\mathbf{A}_1\mathbf{A}_3) \otimes (\mathbf{A}_2\mathbf{A}_4)$. Applying the latter twice, we obtain:

$$\begin{aligned} \mathbf{G}_\mu \mathbf{H}_m \mathbf{C}_m &= (\mathbf{c}_\mu^{\mathcal{H}} \otimes \mathbf{I}_{K+L})(\mathbf{I}_M \otimes \bar{\mathbf{H}}_{m,0} + \mathbf{J}_M \otimes \bar{\mathbf{H}}_{m,1})(\mathbf{c}_m \otimes \mathbf{T}_{zp}) \\ &= (\mathbf{c}_\mu^{\mathcal{H}} \mathbf{c}_m) \otimes (\bar{\mathbf{H}}_{m,0}\mathbf{T}_{zp}) + (\mathbf{c}_\mu^{\mathcal{H}} \mathbf{J} \mathbf{c}_m) \otimes (\bar{\mathbf{H}}_{m,1}\mathbf{T}_{zp}) = \delta(m-\mu)\bar{\mathbf{H}}_{m,0}\mathbf{T}_{zp}, \end{aligned} \tag{6}$$

where we also used the fact that $\bar{\mathbf{H}}_{m,1}\mathbf{T}_{zp} = \mathbf{0}_{(K+L)\times(K+L)}$, which can be readily verified.

Because the last L rows of the matrix $\mathbf{C}_m$ are zero, the IBI is eliminated since $\mathbf{H}_{m,1}\mathbf{C}_m = \mathbf{0}_{P\times K}$. Therefore, using (2) and (6), we can express (3) as:

$$\mathbf{y}_\mu(i) = \mathbf{G}_\mu \mathbf{x}(i) = \bar{\mathbf{H}}_{\mu,0}\mathbf{T}_{zp}\mathbf{s}_\mu(i) + \mathbf{G}_\mu \boldsymbol{\eta}(i). \tag{7}$$

Equation (7) shows how the superposition of received signals from multiple users can be separated deterministically regardless of the unknown FIR multipath channels.

After MUI elimination, any single user equalizer can be applied to $\mathbf{y}_\mu(i)$ of (7) to remove the Inter Symbol Interference (ISI). Since the Toeplitz matrix $\bar{\mathbf{H}}_{m,0}\mathbf{T}_{zp}$ of size $(K+L) \times K$ has *always full rank*, the symbol block $\mathbf{s}_\mu(i)$ is guaranteed to be recoverable. For example, we can adopt the linear ZF equalizer given by († denotes matrix pseudo-inverse) :

$$\boldsymbol{\Gamma}_\mu^{zf} = (\bar{\mathbf{H}}_{\mu,0}\mathbf{T}_{zp})^\dagger, \tag{8}$$

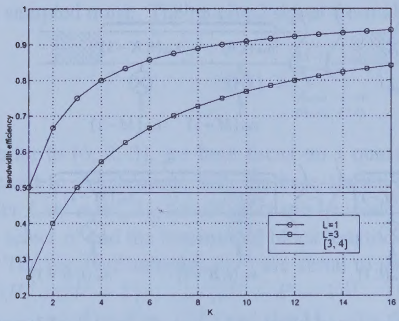


Figure 3: Bandwidth efficiency

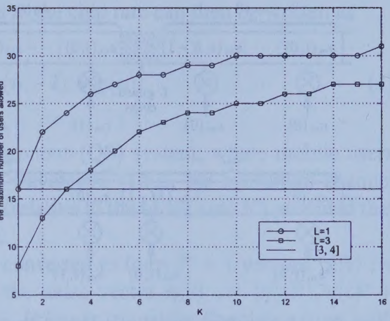


Figure 4: The maximum number of users

or, the linear MMSE, nonlinear DF and ML equalizers.

We next compare the proposed transceiver design with those in [2, 8] and [3, 4], that also achieve deterministic MUI elimination and guarantee symbol recovery; actually, they arrive at the same output [c.f. (7) with [2, eq. (19)] and [4, eq. (9)]].

3.1 Bandwidth Efficiency

Within the received $P \times 1$ block $\mathbf{x}(i)$ in (2), K information symbols are transmitted per user. The bandwidth efficiency for M users can thus be calculated as:

$$\mathcal{E}_1 = MK/[M(K+L)] = K/(K+L). \tag{9}$$

Note that the bandwidth efficiency for the AMOUR system in [2] is: $\mathcal{E}_1' = MK/[M(K+L)+L]$. When $L \ll M(K+L)$, we obtain $\mathcal{E}_1 \approx \mathcal{E}_1'$. On the other hand, choosing $K \gg L$ enables $\mathcal{E}_1$ and $\mathcal{E}_1'$ to approach 1, which indicates that both systems have high bandwidth efficiency. In contrast, the codes in [3, 4] are designed to be shift orthogonal, which constrains their length to equal twice the number of users plus one. Thus, with our notation the bandwidth efficiency in [3, 4] is

$$\mathcal{E}_2 = MK/(2MK+K) = M/(2M+1) < 50\%. \tag{10}$$

Selecting the information block size $K \geq L$ as in [3, 4], we find from (9) and (10) that:

$$\mathcal{E}_1 \approx \mathcal{E}_1' \geq 50\% > \mathcal{E}_2. \tag{11}$$

Another way to look at bandwidth efficiency is to calculate the maximum number of users (with guaranteed MUI elimination and symbol recovery) that can be accommodated by the available system bandwidth. Suppose the system is allocated bandwidth W and the information rate is R_b. The spreading gain is thus $N = W/R_b$. In our system (and [2]), the spreading gain is $P/K = M(K+L)/K$ and thus the maximum number of MUI-free users is

$$M = KN/(K+L), \tag{12}$$

while the maximum number is $(N-1)/2$ for [3, 4] ($N-1$ should also be a power of 2).

To illustrate the difference between our system and [3, 4], we compare in Fig. 3 the corresponding bandwidth efficiencies for channels of maximum order $L = 1$ (2-rays) and $L = 3$ (4-rays) with K varying from 1 to 16. Setting the spreading gain to $N = 33$, Fig. 4 shows the maximum number of MUI-free users that are allowed. Certainly, with increasing K decoding delays increase, but our system can accommodate many more users than [3, 4].

Table 1: complexity of Multiuser Separating front-end

	single user	M users	FFT (or FWT) based algorithms
this paper	$O(M(K+L))$	$O(M^2(K+L))$	$O(M(\log_2 M)(K+L))$
[3, 4]	$O(2M(K+L))$	$O(2M^2(K+L))$	not available
[2, 8]	$O(P(K+L))$	$O(MP(K+L))$	$O(M(\log_2 M)(K+L))$

3.2 Complexity and Design Flexibility

Since MUI-free transceivers decouple MIMO transmissions into multiple parallel SISO transmissions, they have in general lower complexity than joint multiuser detectors (see, e.g., [4] for a comparison). Here we will compare the computational complexity among the three different MUI-resilient transceivers: this paper's and those in [2, 8] and [3, 4]. First, we note that all three transceivers arrive at the same single user block model (7); hence, the complexity of blind channel estimation and equalization is identical. The main difference lies in the multiuser separating front-end matrices $\mathbf{G}_m$. For this reason, we only concentrate on the complexity differences among the three front-ends.

For each user in our proposed transceiver, the front-end consists of $K+L$ inner products between $M \times 1$ vectors. Since each M-long correlator requires M multiplies and $M-1$ adds, its complexity is $O(M)$. So the complexity per user is $O(M(K+L))$ and the overall complexity for all users (e.g., at the Base Station (BS) where we need to demodulate all users' information) is $O(M^2(K+L))$. With M users in [3, 4], the shift-orthogonal codes have length $2M+1$ and subsequently each receiver needs $K+L$ inner products between $2M \times 1$ vectors after discarding 1 cyclic prefix. Therefore, each user in [3, 4] has the complexity of $O(2M(K+L))$, and the overall complexity for M users will be $O(2M^2(K+L))$.

Unlike [3, 4], where only a special class of codes with 50% bandwidth-efficiency is constructed, and different from [2], where highly-efficient codes are generally complex, the design herein is very flexible because the only requirement on our spreading codes is their mutual orthogonality. Since the design of orthogonal codes has been well developed in the literature (at least for multipath-free propagation), there are many fast algorithms available. Using such algorithms, the complexity of our transceivers can be lowered even further. For example, we can adopt IFFT or Walsh-Hadamard (W-H) codes for the code-generating vector $\mathbf{c}_m$ in (4). Then at the BS, we can apply FFT or Fast Walsh Transform (FWT) to separate users. If the FFT is employed, the complexity for all users is $O(M(\log_2 M)(K+L))$, while the FWT is even faster than the FFT.

For each user in [2], the front-end consists of $K+L$ inner products between the $P \times 1$ received block and different Vandermonde vectors. So the complexity per user is $O(P(K+L))$ and the overall complexity for M users will be $O(MP(K+L))$. Note that AMOUR codes can also be constructed based on FFTs to reduce computational complexity. The simplest AMOUR code is proposed in [8, eq. (25)] as:

$$\mathbf{C}_m = \mathbf{f}_m \otimes \mathbf{T}_{zp}, \tag{13}$$

with $\mathbf{f}_m$ the mth column of $M \times M$ FFT matrix. Since $\{\mathbf{f}_m\}_{m=0}^{M-1}$ satisfy d1), it is subsumed in our code designs herein and the resulting receiver has complexity $O(M(\log_2 M)(K+L))$.

For clarity, we summarize complexity requirements and comparisons with [2, 8] and [3, 4] in Table 1. With the same number of users, we infer from Table 1 that our system has less complexity than [3, 4], and the difference becomes more pronounced if special codes (e.g., W-H or FFT) are employed. Fast algorithms for the codes in [3, 4] have not been reported.

A remark is that if channel estimation is complex, or, if the channel is varying fast and channel state information (CSI) needs to be updated frequently, the complexity of the multiuser separating front-end is less instrumental in the overall receiver design.

3.3 Perfectly Constant-Modulus Transmissions

A well-known drawback of multi-carrier systems is that the transmitted signal is not of constant modulus, which imposes restrictions on the system hardware design and undesirable back-offs in the high power amplifier (HPA). Note that the system in [3, 4] has constant modulus, while our system and the FFT-based designs of [2, 8] have constant modulus except at the guard zeros. Although not a problem for the HPA, usage of guard zeros may cause on/off implementation problems in analog system designs. However, with a digital signal processing (DSP) unit and a Digital to Analog Converter (D/A) at the transmitter, inserting zeros is not a problem in e.g., software radio systems. Nevertheless, as we detail next, the proposed transmitter can also dispense with zero padding and achieve perfectly constant modulus transmissions.

With the code design in (4), the $P \times 1$ transmitted block $\mathbf{u}_m(i) = \mathbf{C}_m\mathbf{s}_m(i)$ has M zero sub-blocks of length L evenly distributed across each block. To make up for perfectly constant modulus, we fill in the zero gaps using $L \times 1$ nonzero vectors with entries drawn from the same constellation as $\mathbf{s}_m(i)$; that is, we modify the transmitter as:

$$\check{\mathbf{u}}_m(i) = \mathbf{C}_m\mathbf{s}_m(i) + \check{\mathbf{b}}_m, \quad \check{\mathbf{b}}_m := \mathbf{1}_{M\times 1} \otimes [\mathbf{0}_{K\times 1}^T, \mathbf{b}_m^T]^T \tag{14}$$

where $\mathbf{1}_{M\times 1}$ is the $M \times 1$ vector with unit entries. In this case, the modified transmitted block $\check{\mathbf{u}}_m(i)$ has perfectly C-M across the block provided that the entries of $\mathbf{C}_m$ and $\mathbf{s}_m(i)$ have also C-M.

We next show that these filling vectors do not change our receiver design provided that the codes are selected to have zero mean. With the modified block $\check{\mathbf{u}}_m(i)$ in (14), and the original received vector $\mathbf{x}(i)$ in (2), the new received vector $\check{\mathbf{x}}(i)$ can be expressed as:

$$\check{\mathbf{x}}(i) = \mathbf{x}(i) + \sum_{m=0}^{M-1}(\mathbf{H}_{m,0}\check{\mathbf{b}}_m + \mathbf{H}_{m,1}\check{\mathbf{b}}_m). \tag{15}$$

Thus, our task is to show that $\mathbf{G}_\mu\check{\mathbf{x}}(i) = \mathbf{G}_\mu\mathbf{x}(i)$, which is equivalent to verifying that:

$$\mathbf{G}_\mu(\mathbf{H}_{m,0} + \mathbf{H}_{m,1})\check{\mathbf{b}}_m = \mathbf{0}, \quad \forall m, \mu \in [0, M-1]. \tag{16}$$

Note that $\mathbf{H}_{m,1}$ can be expressed as $\mathbf{J}_M^{-(M-1)} \otimes \bar{\mathbf{H}}_{m,1}$, where $\mathbf{J}_M^{-(M-1)}$ is the $-(M-1)$st power of $\mathbf{J}_M$, which has only a single unity entry at the top-right corner and zeros elsewhere. Introducing the matrix $\mathbf{Z}_M := \mathbf{J}_M + \mathbf{J}_M^{-(M-1)}$ that performs a cyclic shift on a vector and using the block-expression of $\mathbf{H}_{m,0}$ in (5), we then obtain:

$$\mathbf{H}_{m,0} + \mathbf{H}_{m,1} = \mathbf{I}_M \otimes \bar{\mathbf{H}}_{m,0} + \mathbf{Z}_M \otimes \bar{\mathbf{H}}_{m,1}. \tag{17}$$

With the receive matrix $\mathbf{G}_\mu$ designed as (4) and with $\check{\mathbf{b}}_m$ inserted from (14), we arrive at:

$$\begin{aligned}\mathbf{G}_\mu(\mathbf{H}_{m,0}+\mathbf{H}_{m,1})\check{\mathbf{b}}_m &= (\mathbf{c}_\mu^{\mathcal{H}}\mathbf{1}_{M\times 1}) \otimes (\bar{\mathbf{H}}_{m,0}[\mathbf{0}_{K\times 1}^T, \mathbf{b}_m^T]^T) + (\mathbf{c}_\mu^{\mathcal{H}}\mathbf{Z}_M\mathbf{1}_{M\times 1}) \otimes (\bar{\mathbf{H}}_{m,1}[\mathbf{0}_{K\times 1}^T, \mathbf{b}_m^T]^T) \\ &= (\mathbf{c}_\mu^{\mathcal{H}}\mathbf{1}_{M\times 1}) \otimes (\bar{\mathbf{H}}_{m,0} + \bar{\mathbf{H}}_{m,1})[\mathbf{0}_{K\times 1}^T, \mathbf{b}_m^T]^T,\end{aligned} \tag{18}$$

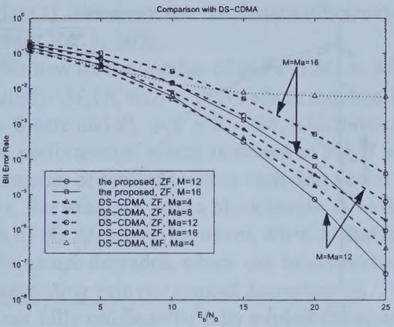


Figure 5: Comparison with DS-CDMA

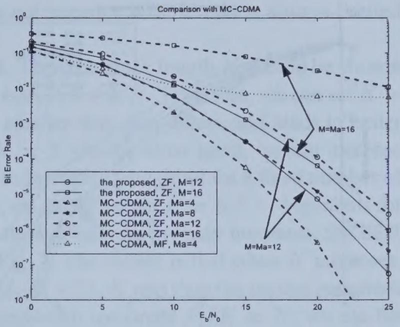


Figure 6: Comparison with MC-CDMA

where the identity $\mathbf{Z}_M \mathbf{1}_{M\times 1} = \mathbf{1}_{M\times 1}$ was used in establishing the second equality.

Therefore, (18) leads to (16) if in addition to d1) we adopt the design constraint of:

d2) balanced user codes; i.e., $\mathbf{c}_\mu^{\mathcal{H}} \mathbf{1}_{M\times 1} = \mathbf{0}$, $\forall \mu \in [0, M-1]$.

Note that d2) is not very restrictive and can be satisfied by many code designs. For instance, IFFT or W-H codes satisfy d2) after discarding only the code with all one entries. The maximum number of users is thus only decreased by one. Recall however that with the filling symbols in (14), we achieve a perfectly constant modulus transmission while maintaining its low-complexity MUI/ISI-resilient reception and its high bandwidth efficiency regardless of frequency-selective multipath.

The penalty with non-zero guards is a small power loss since we allocate $ML/MK = L/K$ percent of the transmit-power for the filling symbols. However, if $K \gg L$, the power loss is small and can be reduced further since the transmitters do not need to pick the same amplitude for the filling symbols as for the data symbols. If the amplitude of the filling symbols is reduced by half, the power loss decreases to one fourth of the original one. The extreme case is to insert filling symbols with zero amplitude as in (4), which incurs no power loss.

4 Simulations

To illustrate the merits of the proposed system, we here present several simulation results.

Test Case 1 (comparison with multiuser detectors): To compare the block-spread MUI-free transceivers with symbol-spread multiuser detectors, we simulate both Direct Sequence (DS) CDMA and Multi Carrier (MC) CDMA systems. The spreading gain is $N = 19$ and the channels have maximum order $L = 3$ (4-rays). Both DS-CDMA and MC-CDMA use W-H spreading codes. When only the channel of the desired user is available, the single-user matched filter (MF) receiver can be employed. The corresponding BER curve levels off and a high error floor appears due to the MUI, as shown in Figs. 5 and 6. To remove MUI, we also simulated ZF multiuser detectors for both DS-CDMA and MC-CDMA, which require knowledge of the codes and the channels of all users at the receiver-end. For the proposed transceiver, we simulated two scenarios. The first uses $K = 6$ to accommodate $M = 12$ users [c.f. (12)]. The second adopts $K = 16$, which enables MUI-free reception of $M = 16$ users. Increasing the block size K allows for more MUI-free users [c.f. (12)]; however, the decoding delay increases and the BER performance per user degrades as illustrated in Figs. 5 and 6 (it will be further elaborated in Te

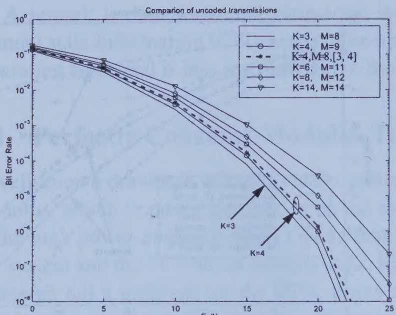


Figure 7: BER comparison with [3, 4]

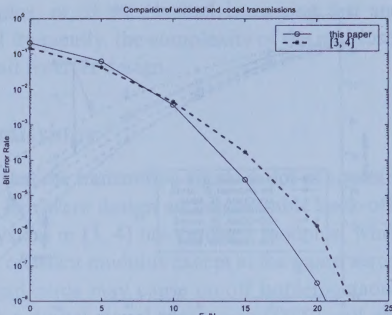


Figure 8: Coding for a fixed number of users

Case 2). Thanks to the MUI-free reception, the BER remains unchanged when the number of active users M_a ($M_a \leq M$) varies. However, for both DS-CDMA and MC-CDMA, when the number of active users M_a increases, the performance of each user degrades, as shown in Figs. 5 and 6. When the system is heavily loaded, e.g., $M_a \approx M$, the proposed MUI-free transceiver outperforms the multiuser-detectors as confirmed by Figs. 5 and 6. When the system is lightly loaded, i.e., $M_a \ll M$, the multiuser detectors could exhibit better performance since MUI is less severe in this case. A final remark here is that multiuser detectors in general require higher computational complexity for both channel estimation and equalization, as illustrated in [4].

Test Case 2 (comparison with the MUI-free transceiver of [3, 4]): To compare the proposed transceiver with those in [3, 4], we here adopt the design parameters of [3] for its shift-orthogonal transceiver: $L = 3$, $K = 4$ and $M_2 = 8$ users. The code length (spreading gain) for each user is then $N = 2M + 1 = 17$ and the block length $P_2 = NK = 68$.

Increasing K does not bring additional benefits to [3, 4]. However, K is very flexible in our system design and different trade-offs can be exploited between decoding delay, BER performance and bandwidth efficiency. To make a fair comparison, we put both transceivers with the same information rate R_b under the same system bandwidth W and thus the same spreading gain $N = W/R_b = 17$. Various system designs can be afforded by the proposed transceiver as detailed in the following:

i) We fix our system to have the same decoding delay by adopting $K = 4$. Therefore, we can afford $M_1 = 9$ users [c.f. (12)], and the block length becomes $P_1 = M_1(K + L) = 63$. Since $M_1 > M_2$ and $P_1 < P_2$, our system accommodates one more user with shorter blocks, which implies that our system has higher bandwidth efficiency than [3]: $\mathcal{E}_1 = M_1K/P_1 > \mathcal{E}_2 = M_2K/P_2$. However, since both systems reach the same block model of (7), each user has the same performance as confirmed by Fig. 7 with $K = 4$. The only difference here is that the system of [3, 4] incurs a small power loss $10\log_{10}(17/16) = 0.26$dB due to the cyclic prefix of length 1 (out of $N = 17$) discarded at the receiver.

ii) As confirmed by (12), the number of users increases for fixed N when K increases, which correspondingly increases the decoding delay. For example, with $K = 6, 8, 14$ we allow for 11, 12, 14 MUI-free users within a block of length $P_1 = 11 \times 9 = 99, 12 \times 11 = 132, 14 \times 17 = 238$, respectively. In addition to longer decoding delays, the BER performance degrades as shown in Fig. 7 as the maximum number of users increases. Clearly, we see the trade-offs among bandwidth efficiency, decoding delays and BER performance. We can also infer from Fig. 7 that increasing K in [3, 4] degrades system performance because it reaches the same model (7). However, the bandwidth efficiency does not improve with K and is always limited to $MK/NK < 50\%$.

iii) We now fix the number of users to $M_1 = M_2 = 8$. Then the block length K should be chosen to satisfy: $M_1(K+L) < KN$, which requires $K > LM_1/(N-M_1)$. Therefore, we can set $K = L = 3$ here and $P_1 = 8 \times 6 = 48$, which results in smaller decoding delays and leads to better BER performance, which is shown in Fig. 7 with $K = 3$. On the other hand, we can increase the number of symbols within each data vector $\mathbf{x}(i)$ to increase throughput for a fixed number of users. For example, with a block length $P_1 = 63, 72$, we can allow for $K = 5, 6 > 4$ symbols for each of the M_1 users. However, when K and hence the symbol rate per user increases, the BER performance degrades, which can be confirmed by Fig. 7. The reason is that when K increases, the spreading gain per symbol decreases: $P_1/K = M_1(K+L)/K$ and thus the system resources are not fully utilized. To fully exploit the system bandwidth resources $P_1/K \approx N$, we can for example transmit the linearly precoded or channel coded $K' \times 1$ $(K' > K)$ vector $\mathrm{s}'_m(i)$, instead of the $K \times 1$ data vector $\mathrm{s}_m(i)$, to improve the BER performance. To illustrate this point, a simple simulation is provided next.

iv) We here simulate $M_1 = M_2 = 6$ users in both systems. Since $N - 1$ should be a power of 2 and $N/2 > 6$ in [3, 4], the code length is again $N = 17$. Here, with the same $K = 4$, and $N = 17$, we use $K' = 7$ and a $(7, 4)$ Hamming code to improve the BER performance, as confirmed by Fig. 8 where the coded BER is calculated through [1, eq. (10.67) and eq. (10.101)]. Both systems also have comparable decoding delays since $K = 4$ data symbols are transmitted over blocks of length $P_1 = 60$ and $P_2 = 68$, respectively.

References

[1] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, Kluwer Academic/ Plenum Publishers, 1999.

[2] G. B. Giannakis, Z. Wang, A. Scaglione, and S. Barbarossa, "AMOUR - Generalized Multicarrier Transceivers for Blind CDMA irrespective of Multipath," *IEEE Trans. on Comm.*, 2000 (to appear); see also *Proc. of GLOBECOM,* 1999, pp. 965-969.

[3] G. Leus and M. Moonen, "MUI-Free Receiver for a Shift-Orthogonal Quasi-Synchronous DS-CDMA System Based on Block Spreading in Frequency-Selective Fading," in *Proc. of Intl. Conf. on ASSP*, June 5-9, 2000, pp. 2497–2500.

[4] G. Leus and M. Moonen, "MUI-Free Receiver for a Synchronous DS-CDMA System Based on Block Spreading in the Presence of Frequency-Selective Fading," *IEEE Trans. on Signal Processing*, 2000 (accepted).

[5] H. Sari and G. Karam, "Orthogonal frequency-division multiple access and its application to catv network," *Euro. Trans. on Telecom.*, pp. 507–516, Nov., 1998.

[6] A. Scaglione, G. B. Giannakis, and S. Barbarossa, "Lagrange/Vandermonde MUI eliminating user codes for quasi-synchronous CDMA in unknown multipath," *IEEE Trans. on Signal Processing*, vol. 48, pp. 2057–2073, July 2000.

[7] S. Verdú, *Multiuser Detection*, Cambridge Press, 1998.

[8] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications: Where Fourier meets Shannon," *IEEE Signal Processing Magazine*, pp. 29–48, May 2000.

Knowing When to Stop

Brendan J. Frey
Department of Computer Science
University of Waterloo
http://www.cs.toronto.edu/~frey

Abstract

In the last decade, simulation has played an important role in demonstrating the excellent performance of iterative decoders. Although techniques for theoretical analysis are now emerging, even the best theoretical papers include simulation results to justify assumptions made in the analysis. A question that arises when running simulations is: "When should the simulation be stopped?" Although the answer to this question has been well-studied in the statistics literature, coding researchers continue to present simulation results without giving readers a sense of how close the results are to the truth. We report the results of an uncontrolled (or maybe, "out of control") experiment carried out on coding researchers during the presentation of this paper at the Allerton conference. The experiment justifies this paper, which reviews some properties of estimators and some simple methods for computing confidence intervals for simulation results.

1 An uncontrolled experiment on coding researchers

Suppose we simulate M blocks in an attempt to estimate the true probability of word error, p_w, of a coding system. Let $e_1, \ldots, e_M$ be indicator variables for word errors. That is $e_i = 1$ indicates that the ith word was erroneous, whereas $e_i = 0$ indicates that the ith word was error-free. A commonly used estimate of p_w is

$$\bar{p}_w = (1/M) \sum_{i=1}^{M} e_i. \tag{1}$$

During the presentation of this paper at the 2000 Allerton conference, the audience was shown the a series of plots from a simulation, showing $\bar{p}_w$ as a function of M for increasing values of M. After being shown each plot, members of the audience were asked to make rough guesses at the true value p_w.

Figs. 1a to 1f show these plots. After seeing only Fig. 1a, the majority of the participants guessed that $p_w < 1 \times 10^{-3}$. After seeing Fig. 1b, the majority guessed

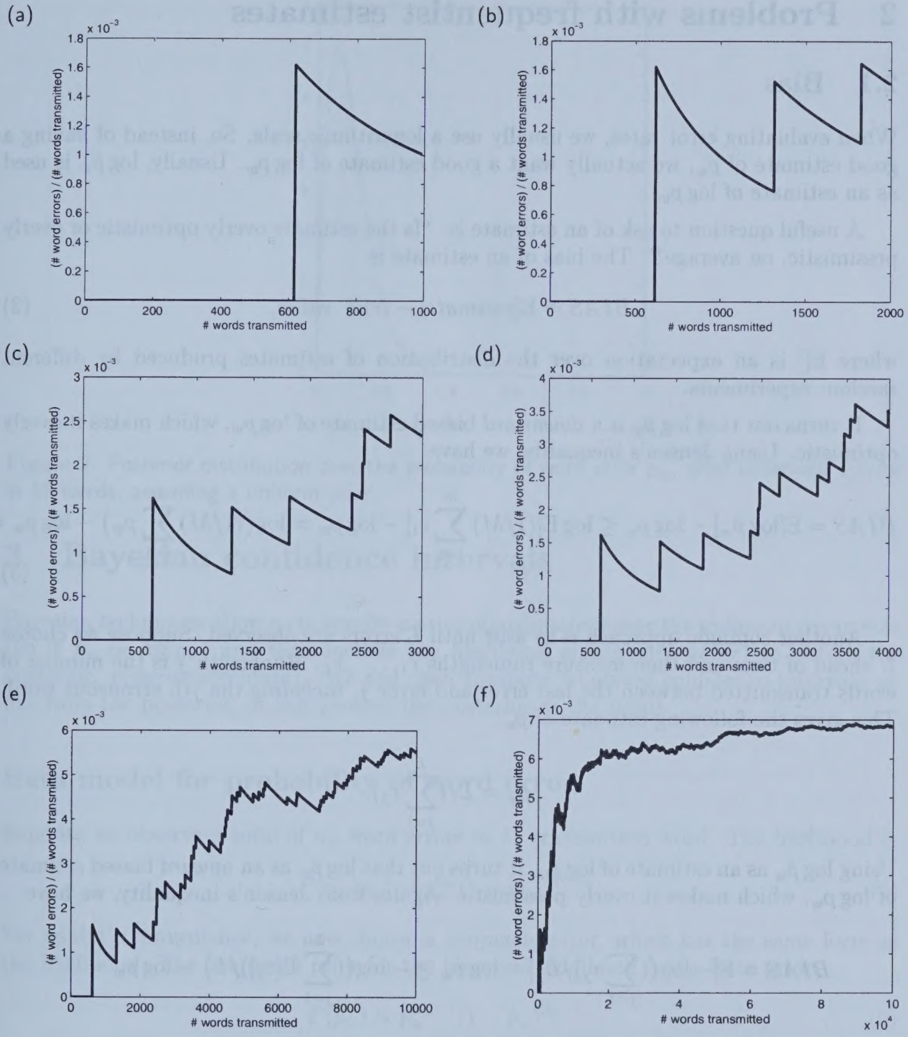


Figure 1: (a) to (f) show the progression of the common estimate of word error rate. In this case, plots (a) to (d) give overly optimistic results. NOTE THE CHANGE IN SCALES.

that $1 \times 10^{-3} < p_w < 1.8 \times 10^{-3}$. After seeing Fig. 1c, the majority guessed that $1.5 \times 10^{-3} < p_w < 3 \times 10^{-3}$. After seeing Fig. 1d, the majority guessed that $p_w < 4 \times 10^{-3}$. Eventually, it was clear that all of these estimates were overly optimistic and that $p_w \approx 7 \times 10^{-3}$ (see Fig. 1f).

Of course, the audience was partly playing into the game of over-optimism. Nonetheless, results *are* frequently reported for simulations where the number of error events puts the results in the category of Figs. 1a to 1d, as opposed to the preferable category of Fig. 1f.

2 Problems with frequentist estimates

2.1 Bias

When evaluating error rates, we usually use a logarithmic scale. So, instead of having a good estimate of p_w, we actually want a good estimate of $\log p_w$. Usually, $\log \bar{p}_w$ is used as an estimate of $\log p_w$.

A useful question to ask of an estimate is: "Is the estimate overly optimistic or overly pessimistic, on average?" The bias of an estimate is

$$BIAS = \mathrm{E}[estimate] - true\ \ value, \tag{2}$$

where E[] is an expectation over the distribution of estimates produced by different random experiments.

It turns out that $\log \bar{p}_w$ is a *downward* biased estimate of $\log p_w$, which makes it overly optimistic. Using Jensen's inequality, we have

$$BIAS = \mathrm{E}[\log \bar{p}_w] - \log p_w \leq \log \mathrm{E}[(1/M)\sum_{i=1}^{M} e_i] - \log p_w = \log\big((1/M)\sum_{i=1}^{M} p_w\big) - \log p_w = 0. \tag{3}$$

Another common approach is to *wait* until L errors are observed. Suppose we choose L ahead of time and then measure runlengths $r_1, \ldots, r_L$. Runlength j is the number of words transmitted between the last error and error j, *including* the jth erroneous word. This gives the following estimate of p_w:

$$\hat{p}_w = L/(\sum_{j=1}^{L} r_j).$$

Using $\log \hat{p}_w$ as an estimate of $\log p_w$, it turns out that $\log \hat{p}_w$ as an *upward* biased estimate of $\log p_w$, which makes it overly pessimistic. Again, from Jensen's inequality, we have

$$BIAS = \mathrm{E}[-\log\big((\sum_{j=1}^{L} r_j)/L\big)] - \log p_w \geq -\log\big((\sum_{j=1}^{L} \mathrm{E}[r_j])/L\big) - \log p_w = 0.$$

2.2 Variance

The variance of $\bar{p}_w$ is easy to estimate:

$$VAR = \mathrm{VAR}[(1/M)\sum_{i=1}^{M} e_i] = \mathrm{VAR}[e_i]/M = p_w(1-p_w)/M. \approx \bar{p}_w(1-\bar{p}_w)/M. \tag{4}$$

However, how can we use an estimate of the variance to produce a confidence interval for $\bar{p}_w$ or $\log \bar{p}_w$? For Gaussian random variables, the variance directly provides, *e.g.*, 95confidence intervals. However, $\bar{p}_w$ is clearly not Gaussian (since its bounded) and $\log \bar{p}_w$ will not generally be Gaussian.

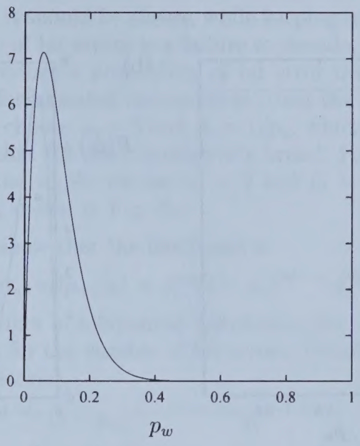


Figure 2: Posterior distribution over the probability of word error p_w, after observing 1 error in 17 words, assuming a uniform prior.

3 Bayesian confidence intervals

Bayesian techniques allow us to specify a prior $p(parameter)$ over the unknown parameter (be it p_w or $\log p_w$) and then include the likelihood $p(data|parameter)$ to obtain the posterior, $p(parameter|data)$. We can then compute whatever confidence intervals we like from the posterior, or just present the posterior as the result.

Beta model for probability of word error

Suppose we observe a total of n_w word errors in M transmitted word. The likelihood is

$$P(n_w|p_w) = p_w^{n_w}(1-p_w)^{M-n_w}. \tag{5}$$

For analytic convenience, we now choose a *conjugate prior*, which has the same form as the likelihood. The conjugate prior for the binomial likelihood is the Beta prior,

$$P(p_w) \propto p_w^{\alpha_w - 1}(1-p_w)^{\beta_w - 1}. \tag{6}$$

Finally, we obtain the posterior,

$$P(p_w|n_w) \propto p_w^{n_w+\alpha_w-1}(1-p_w)^{M-n_w+\beta_w-1}. \tag{7}$$

For example, suppose $M = 18$ and we observe $n_w = 1$. The frequentist estimate is $\bar{p}_e \approx 0.06$ with $2\sqrt{\mathrm{VAR}} \approx 0.1$. Assuming a uniform prior, $\alpha = \beta = 1$, we obtain the posterior, $P(p_w|n_w) \propto p_w(1-p_w)^{17}$. This distribution is shown in Fig. 2.

From this distribution, we can compute whatever confidence intervals we wish to compute. Notice that the confidence intervals will be within the bounds for p_w.

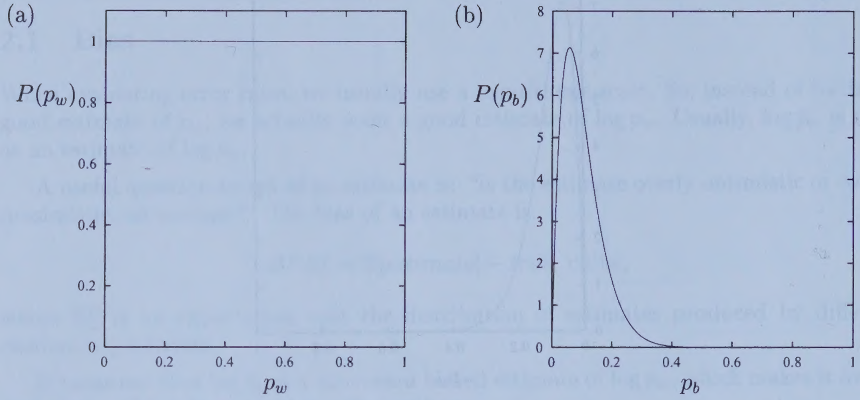


Figure 3: (a) The prior distribution over the probability of word error p_w. (b) The prior distribution over the probability of bit error p_b within erroneous words. This distribution is designed so that its median is equal to the probability of bit error for uncoded transmission.

3.1 Hierarchical Beta model for probability of bit error

In cases where confidence intervals for the probability of bit error are desired, we can use a Beta model for the probability of word error – as described above – and a second Beta model for the probability of bit error within erroneous words.

The error model contains two unknowns: the probability p_w of word error, and the probability p_b of bit error *within erroneous words.* This is a rather crude approximation, since in practice we expect there to be more than one failure mode, *i.e.*, there ought to be several p_b's corresponding to different failure modes.

Let M be the number of words transmitted and let n_w be the number of measured word errors. Let K be the number of information bits per word, and let n_b be the *total* number of bit errors measured while transmitting all M blocks. From the Bayesian perspective, before observing n_w, n_b, we place a prior distribution $P(p_w, p_b)$ on the unknown probabilities. After observing n_w, n_b, we draw conclusions (*e.g.*, compute a confidence interval) from the posterior distribution $P(p_w, p_b|n_w, n_b)$, where

$$P(p_w, p_b|n_w, n_b) \propto P(p_w, p_b)P(n_w, n_b|p_w, p_b). \tag{8}$$

In this equation, the constant of proportionality does not depend on p_w or p_b.

We let p_w and p_b be independent Beta-distributed random variables under the prior: $P(p_w, p_b) = P(p_w)P(p_b)$, where

$$P(p_w) \propto p_w^{\alpha_w - 1}(1-p_w)^{\beta_w - 1}, \quad \text{and} \quad P(p_b) \propto p_b^{\alpha_b - 1}(1-p_b)^{\beta_b - 1}. \tag{9}$$

In frequentist terms, α_w and β_w have the effect of shrinking our measurements toward a word error rate of $\alpha_w/(\alpha_w + \beta_w)$, where the influence of this shrinkage grows with $\alpha_w + \beta_w$. Typically, we choose $\alpha_w = \beta_w = 1$, which gives a uniform prior over p_w as shown in Fig. 3a.

As for the prior over p_b, it should be chosen while keeping in mind the behavior of the decoder. If the main source of bit errors is a failure to decode, and if we believe that for failures the decoder will produce a probability of bit error that is roughly equal to the probability p_u of bit error for uncoded transmission, then the prior should place weight on $p_b = p_u$. In this case, we choose $\alpha_b = 2$ and $\beta_b = 1/p_u$, which ensures that the mode of the prior occurs at p_u and that the prior is relatively broad. For example, for $E_b/N_0 = 1$ dB we have $p_u = 0.0563$, and so We choose $\alpha_b = 2$ and $\beta_b = 1/0.0563 = 17.76$, giving the prior distribution for p_b shown in Fig. 3b.

It is straightforward to show that the likelihood is

$$P(n_w, n_b|p_w, p_b) = P(n_w, n_b|p_w, p_b) \propto p_w^{n_w}(1-p_w)^{M-n_w} p_b^{n_b}(1-p_b)^{n_w K - n_b}. \tag{10}$$

This distribution is the product of a binomial distribution for the number of word errors and a binomial distribution for the number of bit errors. Combining this likelihood with the prior, we obtain the posterior,

$$P(p_w, p_b|n_w, n_b) \propto p_w^{\alpha_w - 1 + n_w}(1-p_w)^{\beta_w - 1 + M - n_w} p_b^{\alpha_b - 1 + n_b}(1-p_b)^{\beta_b - 1 + n_w K - n_b}, \tag{11}$$

which is just the product of a Beta distribution over p_w and a separate Beta distribution over p_b.

Of course, we are actually interested in the posterior distribution $P(p_w p_b|n_w, n_b)$ over the total probability of a bit error $p_w p_b$. We use Monte Carlo to obtain a sample from $P(p_w p_b|n_w, n_b)$. First, we draw p_w — p_b pairs from the posterior in (11) and then we take the product of p_w and p_b in each pair. This sample is sorted in ascending order, and the value of $p_w p_b$ occurring half-way through the sorted list is taken as an estimate of the median of $P(p_w p_b|n_w, n_b)$. Similarly, the values of $p_w p_b$ occurring 2.5% and 97.5% through the sorted list are taken as the 95% confidence interval.

In one of our experiments, we simulated the transmission of $M = 332$ blocks using a block length of $N = 131,072$. We measured $n_w = 14$ and $n_b = 34,225$. Using the prior presented above, a sample of 1000 points from the posterior over p_w and p_b was obtained and is shown in Fig. 4a. As described above, for $\gamma = 0.025$, 0.5 and 0.975, we found the values for p_γ such that $\hat{p}(p_w p_b < p_\gamma|n_w, n_b) = \gamma$, where $\hat{p}$ is the sample distribution. The corresponding three curves of the form $p_w p_b = p_\gamma$ are shown in Fig. 4a, and the corresponding values of p_γ give a median of 1.7×10^{-3} and a 95% confidence interval of $(9.9 \times 10^{-4}, 2.6 \times 10^{-3})$.

Clearly, in this case it is the values for p_w that determine the p_γ's for these curves, whereas the values for p_b are well-determined by the measurements. We could have assumed that p_b took on its measured value instead of sampling from the posterior.

In another experiment, we simulated the transmission of $M = 10,216$ blocks. In this case, we measured $n_w = 0$ and $n_b = 0$. Using naive methods, we might conclude that the bit error rate is 0 and that there isn't any variation in this value. A less naive method would be to assume that the *next* block would be in error and pick a value for n_b.

The Bayesian technique gives the sample from the posterior shown in Fig. 4b. In this case, the values of both p_w and p_b play a role in determining the p_γ's for the three curves. The median is 5.1×10^{-6} and the confidence interval is $(1.6 \times 10^{-7}, 4.8 \times 10^{-5})$.

Summary

We illustrated that just "watching" simulations can lead a researcher astray when trying to obtain an accurate estimate of word error rate or bit error rate. Frequentist estimates

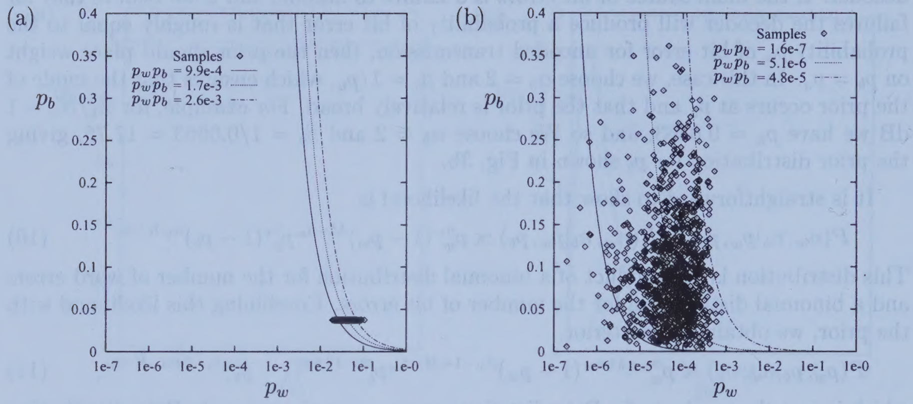


Figure 4: (a) A 1000-point sample from $P(p_w, p_b|n_w, n_b)$ for $M = 332$, $n_w = 14$, $K = 65{,}536$ and $n_b = 34{,}225$, for the prior described in the main text. (b) A 1000-point sample from $P(p_w, p_b|n_w, n_b)$ for $M = 10{,}216$, $n_w = 0$, $K = 65{,}536$ and $n_b = 0$, for the same prior.

can be biased upward or biased downward and there is little control over this bias. Further, frequentist confidence intervals may not make sense for bounded parameters.

Bayesian techniques combine a user-specified prior over the unknown parameter with the observation likelihood, producing a posterior distribution over the parameter of interest. This posterior can be used to compute a confidence interval. Conjugate priors can be used to simplify analysis, but if there is a good reason to use another type of prior, Monte Carlo methods can be used to obtain a sample from the posterior. Inferences can then be made using the sample.

References

Your undergraduate textbook on probability and statistics. Read it, and put confidence intervals on your plots from now on!

Soft-Input/Soft-Output Sequential Decoding[*]

Gennaro Lanzetta
Dept. of Electrical Engineering
Swiss Federal Institute of Technology
8092 Zurich, Switzerland
glanzett@ee.ethz.ch

Hermano A. Cabral   Daniel J. Costello, Jr.
Dept. of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
{hcabral,dcostell1}@nd.edu

Abstract

A soft-input/soft-output sequential decoder (SISOSD) is proposed along with an efficient method for the extraction of soft-outputs. Subsequently, the SISOSD is embedded in an iterative decoding scheme for decoding parallel concatenated long constraint length convolutional codes.

1 Introduction

Recent research[1] indicates that long constraint length (CL) recursive convolutional codes may be of interest as component codes in Turbo decoding, if they can be effectively decoded. This immediately suggests that a soft-input/soft-output sequential decoder (SISOSD) may be useful in a Turbo coding system. A SISOSD would allow a large free distance convolutional code to be used as a component code and thereby eliminate the characteristic error-floor seen with Turbo codes at moderate to high signal-to-noise ratios. Finally, a reduction in computational effort per decoded bit compared to the well-known BCJR algorithm used with MAP decoding is expected.

Previous attempts towards a SISOSD include the twin stack algorithm[2]. That approach, however, does not extend to long CL codes, and therefore it still suffers from an error floor. This paper begins with the development of a modified Fano metric suitable for including *a priori* information about the source bit probability. It also introduces an efficient method for the extraction of soft-outputs from the sequential decoding tree, followed by simulation results of a turbo-like scheme using the SISOSD.

2 The SISOSD and Its Applications

Consider a source that generates the information sequence $\boldsymbol{u} = \boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{L-1}$ of k-tuples $\boldsymbol{u}_j = (u_j^1, \ldots, u_j^k)$, $u_j^i \in \{-1, +1\}$, to be encoded by a rate $R = k/n$ systematic recursive convolutional encoder into the code sequence $\boldsymbol{x} = \boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{L-1}$ of n-tuples $\boldsymbol{x}_j = (x_j^1, \ldots, x_j^n)$, $x_j^i \in \{-1, +1\}$, for transmission over an AWGN channel, producing the received sequence $\boldsymbol{y} = \boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{L-1}$ of n-tuples $\boldsymbol{y}_j = (y_j^1, \ldots, y_j^n)$, $y_j^i \in \mathbb{R}$.

By revisiting the variable-length decoding problem[3], the Fano metric can be modified to incorporate *a priori* information about the source and to take into account the systematic structure of the code. Using the fact that the information k-tuple $\boldsymbol{u}_j$ is contained in the encoded n-tuple $\boldsymbol{x}_j = (\boldsymbol{u}_j, x_j^{k+1}, \ldots, x_j^n)$, it follows that

$$\mu_{\mathrm{SYS}}(\boldsymbol{x}) = \sum_{j=0}^{L-1} \left\{ \sum_{i=1}^{k} \left(\log \left(\frac{P(y_j^i | u_j^i)}{P_0(y_j^i)} \right) + \log P(u_j^i) \right) + \sum_{i=k+1}^{n} \log \left(\frac{P(y_j^i | x_j^i)}{P(y_j^i)} \right) \right\},$$

---

[*] This work was supported by NASA Grant NAG5-8355 and NSF Grant CCR00-75514.

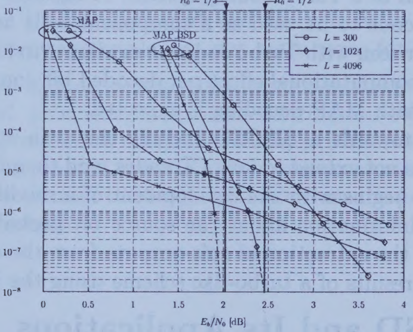
where $P_0(y_j^i) = P(u_j^i = -1)P(y_j^i|u_j^i = -1) + P(u_j^i = +1)P(y_j^i|u_j^i = +1)$. The usefulness of *a priori* information in sequential decoding was previously shown by means of extensive computer simulation in [4].

The generation of soft-outputs is based on post-processing, i.e., the soft-outputs are extracted after the sequential decoding procedure has terminated or after a buffer overflow. The starting point is the decoding tree composed of all nodes that have been examined (visited) by the sequential decoding algorithm. The visited nodes are labeled with their computational loads, i.e., the number of nodes contained in the subtree stemming from that node. To keep the notation simple, we will assume $k = 1$, but the generalization is straightforward. Letting $C_{u_j,j}$ denote the sum of the computational loads of all nodes at level j corresponding to the source bit u_j, then the *a posteriori* probability of u_j is approximated by

$$P(u_j|\boldsymbol{y}) \approx \frac{C_{u_j,j}}{C_{0,j} + C_{1,j}},$$

from which the extrinsic information can be extracted in a similar way to Turbo decoding.

As an application of the SISOSD, a rate 1/3 Turbo code can employ the SISOSD to decode a long CL code used as component code. The best performance, shown below, was obtained in a hybrid scheme where only one of the component codes is a long CL code and a bidirectional sequential decoder (BSD) [5] is used in the SISOSD.



References

[1] P.C. Massey and D.J. Costello, Jr.: "New developments in asymmetric turbo codes," in Proc. Int. Symposium on Turbo Codes, Brest, France, Sept. 2000.

[2] R. Sivasankaran and S.W. McLaughlin: "Twin-stack decoding of recursive systematic convolutional codes," in Proc. 37th Allerton Conf., Monticello, IL, Oct. 1999.

[3] J. L. Massey: "Variable-length codes and the Fano meric," *IEEE Trans. Inf. Theory*, vol.IT-18, no.1, pp.196–198, Jan. 1972.

[4] C. Weiss, S. Reidel, and J. Hagenauer: "Sequential decoding using *a priori* information," *Electronic Letters*, vol.32, no.13 pp.1190–1191, June 20, 1996.

[5] S. Kallel and K. Li: "Bidirectional sequential decoding," *IEEE Trans. Inf. Theory*, vol.IT-43, no.4, pp.1319–1326, July 1997.

Recursive constructions and their maximum likelihood decoding

Ilya Dumer and Kirill Shabunov*

Abstract

We consider recursive decoding techniques for RM codes, their subcodes, and newly designed codes. For moderate lengths up to 512, we obtain near-optimum decoding with feasible complexity.

1 Introduction

In this paper, we consider decoding algorithms that can achieve good performance and low complexity on moderate blocklengths. Our goal is to fill the void left by the best algorithms, such as optimum maximum likelihood (ML) decoding, which has unfeasible complexity even on relatively short blocks, and iterative decoding, which becomes very efficient beginning with the lengths of tens of thousands. More specifically, we wish to achieve near-optimum performance on the lengths ranging from 128 to 512, where neither of these two algorithms can yet combine good performance with low complexity.

To achieve this goal, we will use *recursive techniques.* One particular class of codes generated by (multilevel) recursion is Reed-Muller (RM) codes and their subcodes. Also, RM codes are only slightly inferior to the best codes on moderate lengths. We will see below that recursive decoding substantially outperforms other (nonexponenential) algorithms known for RM codes. Our basic recursive procedure will split the *RM* code (r, m) of length n into the two constituent RM codes $(m\text{-}1, r\text{-}1)$ and $(m\text{-}1, r)$ of length $n/2$. Decoding is then relegated further to the shorter codes until we reach basic codes with feasible ML decoding. In all intermediate steps, we only recalculate the reliabilities of the newly defined symbols.

To improve decoding performance, we will also generalize recursive design. In particular, we use subcodes of RM codes and their modifications. We also use relatively short lists of code candidates in the intermediate steps of the recursion. As a result, we closely approach ML decoding performance on the blocklengths up to 512.

2 Reed-Muller codes

We use notation $\left\{ {m \atop r} \right\}$ for RM codes (n, k) of length $n = 2^m$, dimension $k = \sum_{i=0}^{r} \binom{m}{i}$ and distance $d = 2^{m-r}$. RM codes found numerous applications thanks to fast decoding procedures. First, *majority algorithm* [6] enables feasible bounded-distance decoding and can even correct [3] most error patterns of weight up to $(d \ln d)/4$ on long codes of fixed rate R.

---

*The authors are with the College of Engineering, University of California, Riverside, CA 92521. This research was supported by the NSF grant NCR-9703844.

Other efficient decoding schemes are based on recursive technique of [5] and [2]. These algorithms enable bounded distance decoding with the lowest complexity order of $n \min(r, m-r)$ known for RM codes. Simulation results [8] show that recursive algorithms increase decoding domain of bounded distance decoding. Subsequently, these algorithms were slightly refined in [9]. It was shown that (similar to majority decoding) recursive algorithms of [5] and [2] correct most error patterns up to the weight $(d \ln d)/4$ when used on long codes of fixed rate R.

For long low-rate RM codes of fixed order r, both majority decoding and recursive schemes correct most error patterns of Hamming weight up to $n(1-\varepsilon_r^{\mathrm{maj}})/2$, where the residual term has vanishing order

$$\varepsilon_r^{\mathrm{maj}} \sim (m/d)^{1/2^{r+1}} \tag{1}$$

as $m \to \infty$. Note that (1) gives a *threshold-type* capacity that approaches the upper limit of $n/2$. However, degree of convergence is relatively slow even for codes $\left\{{m \atop 2}\right\}$. Much better results are obtained for ML decoding. For long codes $\left\{{m \atop r}\right\}$ of fixed order r, it is proven in [7] that ML decoding further reduces the residual term $\varepsilon_r^{\mathrm{maj}}$ to the order of

$$\varepsilon_r^{\mathrm{ML}} \precsim (\ln 4)\sqrt{m^r/n}. \tag{2}$$

3 Recursive structure

In essence, all recursive techniques known for RM codes are based on the *Plotkin construction.* Here the original RM code $\left\{{m \atop r}\right\}$ is represented in the form $(u,u+v)$, by taking any subblock u from RM $\left\{{m-1 \atop r}\right\}$ and any v from RM $\left\{{m-1 \atop r-1}\right\}$. These two subcodes have length 2^{m-1}. By continuing this process, we again obtain the shorter RM codes of length 2^{m-2} and so on. Finally, we arrive at the end nodes that are repetition codes $\left\{{j \atop 0}\right\}$ and full spaces $\left\{{j \atop j}\right\}$. This is schematically shown in Fig. 1 for RM codes of length 32. In Fig. 2, we consider incomplete decomposition terminated at the biorthogonal codes $\left\{{j \atop 1}\right\}$ and single-parity check codes $\left\{{j \atop j-1}\right\}$.

Now let I_r^m denote a block of information bits that encodes a vector $(u, u+v)$. It is also important that our recursion splits I_r^m into two information subblocks I_r^{m-1} and I_{r-1}^{m-1} that encode vectors u and v, respectively. Correspondingly, code dimensions satisfy the recursion $|I_r^m| = |I_r^{m-1}| + |I_{r-1}^{m-1}|$. In this way, the shorter information subbloks can be split again until we arrive at the end nodes. Thus, any specific codeword can be encoded from the (multiple) information strings assigned to the end nodes $\left\{{j \atop 0}\right\}$ or $\left\{{j \atop j}\right\}$. Following [2], it can be proven that recursive encoding of code $\left\{{m \atop r}\right\}$ has complexity

$$\psi_r^m \leq n \min(r, m-r) + 1. \tag{3}$$

This observation comes from two facts. First, the end nodes $\left\{{j \atop 0}\right\}$ and $\left\{{j \atop j}\right\}$ satisfy the bound (3). Second, consider the two constituent codes $\left\{{m-1 \atop r-1}\right\}$ and $\left\{{m-1 \atop r}\right\}$. Then $(u, u+v)$ construction gives complexity $\psi_{r-1}^{m-1} + \psi_r^{m-1} + \frac{n}{2}$ for the code $\left\{{m \atop r}\right\}$. Using this recursion, one can show that ψ_r^m also satisfies (3) if constituent codes do.

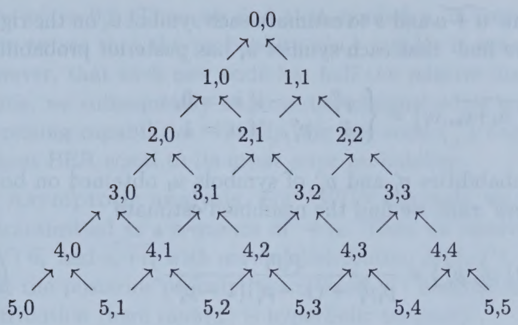


Fig. 1: Full decomposition

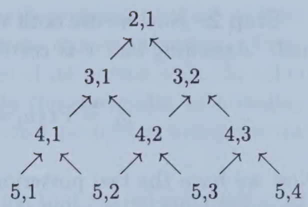


Fig. 2: Partial decomposition

4 New decoding techniques

Our algorithm also uses the $(u,u+v)$ construction and relegates decoding to the two constituent RM codes. Decoder receives a block $(\widetilde{u}, \widetilde{u+v})$ that consists of two halves $\widetilde{u}$ and $\widetilde{u+v}$ corrupted by noise. We first try to find the better protected codeword v from $\left\{ {m-1 \atop r-1} \right\}$. Then we proceed with the block u from the code $\left\{ {m-1 \atop r} \right\}$. In a more general scheme, we repeat this recursion, by decomposing subblocks v and u further. On all intermediate steps, we only recalculate the probabilities of the newly defined symbols. Finally, we perform soft decision ML decoding once we reach the end nodes. The most important difference from the previous work [9] is that in each step we keep L most probable candidates obtained prior to this step. This difference is discussed in Section 5. In this section, we first assume that our decoding is terminated on the biorthogonal codes depicted in Fig. 2.

Step 1. To find a subblock v in hard-decision decoding, one would use its corrupted version $\widetilde{v} = \widetilde{u} + \widetilde{u+v}$. Using more general approach, we find the posterior probabilities of the received symbols. On the left half $\widetilde{u}$, each symbol u_i has posterior probability

$$p'_i \stackrel{\text{def}}{=} \Pr\{u_i = 0 \,|\widetilde{u}_i\}.$$

Similarly, we use the right half $\widetilde{u+v}$ to find the posterior probability of any symbol $u_i + v_i$:

$$p''_i \stackrel{\text{def}}{=} \Pr\{u_i+v_i = 0 \mid \widetilde{u_i+v_i}\}.$$

Given the probabilities p'_i and p''_i of the symbols u_i and u_i+v_i, we then find the posterior probability $p(v_i)$ of their binary sum v_i. Here we use the formula of total probability and find

$$p(v_i) \stackrel{\text{def}}{=} \Pr\{v_i = 0 \mid \widetilde{u}_i, \widetilde{u_i+v_i}\} = p'_i p''_i + (1-p'_i)(1-p''_i). \qquad (4)$$

Here we use the fact that the two original symbols u_i and u_i+v_i are independent. Also, both symbols are independently corrupted by Gaussian noise. Now we can use any soft-decision decoding that uses probabilities $p(v_i)$ to find the most probable vector v from the $\left\{ {m-1 \atop r-1} \right\}$-code. This completes Step 1 of our algorithm. Vector v is then passed to Step 2.

Step 2. Now we use both vectors $\widetilde{u+v}$ and v to estimate each symbol u_i on the right half. *Assuming that v is correct,* we find that each symbol u_i has posterior probability

$$p_i^{\wedge} \overset{\text{def}}{=} \Pr\{u_i = 0 \mid \widetilde{u_i{+}v_i},\, v_i\} = \begin{cases} p_i'', & \text{if } v_i = 0, \\ 1 - p_i'', & \text{if } v_i = 1. \end{cases}$$

Now we have the two posterior probabilities p_i' and $p_i^{\wedge}$ of symbols u_i obtained on both corrupted halves. By using the Bayes' rule, we find the combined estimate

$$p(u_i) \overset{\text{def}}{=} \Pr\{u_i = 0 \mid p_i', p_i^{\wedge}\} = \frac{p_i' p_i^{\wedge}}{p_i' p_i^{\wedge} + (1 - p_i')(1 - p_i^{\wedge})}. \tag{5}$$

Finally, we perform soft decision decoding and find a subblock $u \in \left\{ {m-1 \atop r} \right\}$.

Thus, procedure $\left\{ {m \atop r} \right\}$ has a recursive structure that calls procedures $\left\{ {m-1 \atop r-1} \right\}$ and $\left\{ {m-1 \atop r} \right\}$, and so on. By recalculating probabilities (4) and (5), we finally arrive at the *biorthogonal* Reed-Muller codes $\left\{ {j \atop 1} \right\}$ on our way to the left, or full codes $\left\{ {j \atop j} \right\}$ on the way to the right. *Maximum likelihood* decoding is executed on the end nodes. Each decoding retrieves a new subset of information symbols associated with the current end node. In both cases, maximum likelihood decoding has complexity order at most $n \log_2 n$ [4]. Simple analysis also shows that recalculating all posterior probabilities in (4) and (5) has complexity at most $5n$. Therefore our decoding complexity Ψ_r^m satisfies the recursion

$$\Psi_r^m \le \Psi_{r-1}^{m-1} + \Psi_r^{m-1} + 5n.$$

This brings the overall complexity to the order of $5n \log_2 n$ real operations. A slightly more efficient version gives complexity $(5n \log_2 n)/2$.

5 Analysis and improvements

Given the code $\left\{ {m \atop r} \right\}$, we first decode code $\left\{ {m-r+1 \atop 1} \right\}$ followed by codes $\left\{ {m-r \atop 1} \right\}, \left\{ {m-r-1 \atop 1} \right\}$, and so on. With the exception of the leftmost and the rightmost nodes, the procedure enters each node multiple times, by taking all the paths leading to this node. It turns out that the output bit error rate (BER) significantly varies on different nodes and even on different paths leading to the same node. Therefore our first problem is to define the most error-prone paths. We start our analysis with two examples.

Example 1. For simplicity, assume that the all-zero codeword from the code $\left\{ {m \atop r} \right\}$ is transmitted over the binary channel with crossover probability $p = 0.9$. Then we use formula (4) with $p_i' = p_i'' = 0.9$ to find the probability $p(v_i)$ of correct symbol $v_i = 0$ in the block $v \in \left\{ {m-1 \atop r-1} \right\}$. From (4) we see that $p(v_i) = 0.82$. Subsequently, this probability $p(v_i)$ rapidly converges to 0.5 in a few more steps. On the positive side, we note that each step gives us a better protected code that has twice the relative distance of the former one. In particular, the leftmost node $\left\{ {m-r+1 \atop 1} \right\}$ has length $2d$ and distance d. Its ML decoding gives asymptotically vanishing BER if the residual term $\varepsilon_1^{\mathrm{ML}}$ still exceeds $(\ln 4)\sqrt{m/2d}$, according to (2).

Example 2. Suppose that our original code $\left\{ {m \atop r} \right\}$ from the previous example has already received the correct subblock v from the code $\left\{ {m-1 \atop r-1} \right\}$. Now we need to find the remaining subblock u from the code $\left\{ {m-1 \atop r} \right\}$. Given correct v, we can use (5) with

$p'_i = p^{\wedge}_i = 0.9$. Then we find that symbol u_i is correct with probability $p(u_i) \approx 0.99$. Now we see that the probability $p(u_i)$ rapidly increases as we move to the right. Note, however, that each new code has half the relative distance of its parent code. In other words, we subsequently improve the channel while entering the new codes with weaker correcting capabilities. Finally, the last code $\left\{ {r \atop r} \right\}$ has no error protection and gives the output BER equal to its input error probability.

Asymptotic analysis. For AWGN channels, we assume that the all-zero codeword is transmitted as a sequence of $+1$s. Then we receive n independent random variables (RV) $\widetilde{u}_i$ and $\widetilde{u_i{+}v_i}$ with normal distribution $\mathcal{N}(1, \sigma^2)$. Accordingly, it can be readily seen that the posterior probabilities p'_i (and p''_i) become independent RV with non-Gaussian distribution (here $\tanh(x)$ is hyperbolic tangent):

$$p'_i = (1 + \varepsilon'_i)/2, \qquad \text{where } \varepsilon'_i = \tanh(2\widetilde{u}_i/\sigma^2). \tag{6}$$

In the next step, we obtain the RV $p(v_i)$ and $p(u_i)$. Their distributions can also be written (see [9]) in the form (6), where the residual terms are:

$$\varepsilon_i(v_i) = \varepsilon'_i \varepsilon''_i, \qquad \varepsilon_i(u_i) = \tanh(2\widetilde{u}_i/\sigma^2 + (-1)^{v_i} 2(\widetilde{u_i{+}v_i})/\sigma^2). \tag{7}$$

Similar to Example 1, it can be shown that the first RV $\varepsilon_i(v_i)$ has a smaller expectation $\overline{\varepsilon_i(v_i)}$ relative to the original estimate ε'_i. By contrast, the second RV $\varepsilon_i(u_i)$ has a greater expected value. Here the analysis is similar to Example 2. We also use that the newly defined RV $\varepsilon_i(v_i)$ and $\varepsilon_i(u_i)$ are all independent for each new step. Now consider asymptotic case of high noise power $\sigma^2 \gg 1$. (Note that this case is relevant to long RM codes with $m \to \infty$ and fixed order r.) Then we use asymptotic approximations in formulas (4) and (5) and arrive at the following conclusions.

- We prove that moving to the left from $\left\{ {m \atop r} \right\}$ to $\left\{ {m-1 \atop r-1} \right\}$ and further is equivalent to squaring our noise power (bringing it to σ^4, then σ^8, and so on), while keeping the signal energy equal to 1. By contrast, the original noise power σ^2 is cut by half when the algorithm moves to the right (bringing it to $\sigma^2/2$, $\sigma^2/4$, and so on).
- We prove that the left-hand movement makes our subcodes much more vulnerable. In this case, doubling the relative code distance d/n does not compensate for a stronger noise. In particular, the highest (worst) BER $\mathcal{P}_1$ is obtained on the leftmost node $\left\{ {m-r+1 \atop 1} \right\}$ that is decoded first. The second worst BER $\mathcal{P}_2$ is obtained on the next decoded node $\left\{ {m-r \atop 1} \right\}$, and so on. Using conventional notation $Q(x) = \int_x^\infty e^{-u^2/2} du/\sqrt{2\pi}$, we prove that for $m \to \infty$:

$$\mathcal{P}_1 \sim Q(2^{(m-r)/2} \sigma^{-2^{r-1}}), \qquad \mathcal{P}_2 \sim Q(2^{(m-r+1)/2} \sigma^{-2^{r-1}}). \tag{8}$$

Now we see that even two adjacent nodes give very different results, where $\mathcal{P}_2 \sim \mathcal{P}_1^2$ for small $\mathcal{P}_1$. By contrast, moving to the right does not increase the output BER relative to the parent code. In this case, the lowest BER is obtained on the rightmost node $\left\{ {r \atop r} \right\}$.

Asymptotic comparison. For long RM codes new recursive decoding increasingly outperforms both the majority algorithm and the former recursive techniques of [5], [8] as the block length grows. In particular, these algorithms give BER $\mathcal{P} \sim Q(2^{(m-r)/2} \sigma^{-2^r})$. Further, it can be shown that for long RM codes of fixed rate R, the above decoding corrects most error patterns of weight up to $(d \ln d)/2$ thus:

- increasing $\ln d$ times the capacity of bounded-distance decoding;
- doubling the capacity $(d \ln d)/4$ of the former recursive technique.

Improvements. An important conclusion resulting from the above analysis is to set the leftmost information bits as zeros. In this way, we arrive at the subcodes of the original code $\genfrac{\{}{\}}{0pt}{}{m}{r}$ that are obtained by eliminating only a few least protected information bits. In particular, even eliminating the first $m-r+2$ information bits that form the leftmost code $\genfrac{\{}{\}}{0pt}{}{m-r+1}{1}$, immediately can reduce the output BER from $\mathcal{P}_1$ to its square $\mathcal{P}_2$ for sufficiently long codes.

Decoding performance can be further improved by using *list decoding.* To simplify the analysis, we now consider the repetition codes $\genfrac{\{}{\}}{0pt}{}{j}{0}$. In particular, we start with the leftmost code $\genfrac{\{}{\}}{0pt}{}{m-r}{0}$ and take *both* codewords $v = 0$ and $\overline{v} = 1$ of length 2^{m-r}. Correspondingly, we keep both posterior probabilities instead of choosing the more probable codeword. This step gives the two initial edges of a tree. Each edge is associated with a cost function equal to the log of the (corresponding) posterior probability.

Then we decode the next code $\genfrac{\{}{\}}{0pt}{}{m-r-1}{0}$. Note that the former codewords v and $\overline{v}$ give different probability distributions on this node. Given v and $\overline{v}$, our new decoding is performed 2 times, separately for v and $\overline{v}$. The result is a full tree of depth 2, that has 4 new edges along with their cost functions. The next step includes 4 decodings of the code $\genfrac{\{}{\}}{0pt}{}{m-r-2}{0}$ performed on each path of the tree. By continuing this process, we arrive at the codes $\genfrac{\{}{\}}{0pt}{}{1}{0}$ and $\genfrac{\{}{\}}{0pt}{}{1}{1}$. We also keep accumulating the posterior probabilities of our paths. It can be seen that the resulting 2^{m-r+2} paths give full biorthogonal code $\genfrac{\{}{\}}{0pt}{}{m-r+1}{1}$. Choosing the best path at this point becomes equivalent to the original termination at the biorthogonal codes.

To improve our decoding, we keep all L paths instead of selecting the best paths. In a more general scheme, the threshold L can be greater or smaller than 2^{m-r+2}. In any case, we start at the repetition codes and keep doubling[1] the number of paths until $2L$ paths are formed. After $2L$ paths are constructed, we choose L paths with L maximum cost functions. In the end, the most probable path (that is, the path with the maximum cost function) is chosen among L paths survived at the rightmost node.

Both the simulation results and calculations show that continuous regeneration of L best candidates improves our original algorithm that selected the best path at each node. In other words, keeping the longer paths allows us to better separate the transmitted vector from the remaining candidates. As a result, we substantially reduce the overall BER even when compared to the expurgated subcodes. Note, however, that our list decoding increases complexity L times, to the order of $Ln \log_2 n$. To refine this scheme further, recall that the channel quality constantly improves as we move from the left to the right. Therefore, we can choose the variable threshold L that becomes smaller as our decoding progresses to the rightmost nodes. In this way, we can substantially reduce our list-decoding complexity even when L originally exceeds n.

Simulation results. Our results are described below in Figures 4 to 9. These figures also reflect the drastic improvements obtained when both techniques - using the subcodes and short decoding lists - were combined. The curves with $L = 1$ show the performance of the refined version of the former recursive techniques from [5], [2], and [8]. For codes of length 256 and 512, the results are now improved by 3.5 to 5 dB at BER 10^{-4}.

While using the maximum lists depicted on each figure, simulation also showed that in most cases of incorrect decoding, the erroneous result is more probable than the transmitted vector. This fact shows that our block ER (BL ER) is very close to that of ML decoding. In turn, this gives a new (experimental) bound on the BL ER of ML decoding. Also, our results substantially surpass other codes with similar parameters (see the current "world records" on http://www331.jpl.nasa.gov/). In Fig. 9, we summarize the results on block ER of ML decoding for RM codes $\left\{ {8 \atop 2} \right\}$ to $\left\{ {8 \atop 6} \right\}$ of length 256.

---

[1] We can also increase the number of paths, say, to $4L$ or $8L$ on the nodes RM (j, j).

It is also interesting that subcodes usually achieve near-ML decoding using much smaller lists relative to the original RM codes. In particular, a subcode (256,78) approaches near-ML decoding using *only 32 intermediate paths.* Note that even one of the most efficient algorithms developed in [1] uses about 10^5 paths for BCH codes of length 256. On the other hand, our simulation results show that codes of length 512 approach ML decoding using much bigger lists than codes of length 256. To extend the results for longer codes, we use slightly different constructions described in the next section.

6 More general recursive constructions

Multiple splitting of RM codes. Here we wish to change the original Plotkin representation. Namely, one can apply more sophisticated partitions that directly split RM codes in 4, 8, or more codes of shorter lengths. For example, by applying Plotkin construction two times, we can split the original block into four quarters $(u, u + w_1, u + w_2, u + w_1 + w_2 + v)$. Here u is taken from the least protected code $\left\{ {m-2 \atop r} \right\}$, vectors w_1 and w_2 belong to the medium-protected code $\left\{ {m-2 \atop r-1} \right\}$, while v is taken from the best protected code $\left\{ {m-2 \atop r-2} \right\}$. Simulation performed for this construction did not improve the results presented in Figures 4 to 9.

Slightly better results were obtained for low SNR, when these four codes were combined in a different way as $(u, u + w_1, u + w_1 + w_2, u + w_1 + w_2 + v)$. It can be proven that for low rates the latter construction gives asymptotic improvement to our original Plotkin representation. This conclusion stems from the following facts. As before, the code v is decoded first and is most vulnerable in recursive decoding. Note also that v is obtained directly in one step, by adding the third quarter $u + w_1 + w_2$ and the forth quarter $u + w_1 + w_2 + v$ of our construction. Asymptotically, such a step squares the noise power, as described above. On the other hand, we reduce the length four times in each step. Accordingly, the new recursive construction reaches the leftmost nodes $\left\{ {j \atop 0} \right\}$ in $r/2$ steps instead of r steps used before. As a result, we can replace the former term $\sigma^{-2^{r-1}}$ in (8) by the greater term $\sigma^{-2^{r/2}}$.

Despite substantial asymptotic improvements, simulation showed that these improvements start accumulating only on the lengths of 2048 and above.

Alternating recursions. Suppose that we use the Plotkin construction $(u, u + v)$ in Fig. 1, but change our original code v from $\left\{ {m-1 \atop r-1} \right\}$ to $\left\{ {m-1 \atop r-2} \right\}$. In other words, we move one more step to the left relative to the Plotkin construction as shown in Fig. 3. As a result, the new code v has a better error protection. This alteration also doubles the distance of code v and gives unequal error protection for the original code. On the other hand, we also reduce the overall code rate and the SNR per channel symbol (given the same SNR per information bit). This lower rate can eliminate the advantages of the better protection. To increase code rate in v, we then add extra symbols in the next splitting step. For example, we split v into codes $\left\{ {m-2 \atop r-3} \right\}$ and $\left\{ {m-2 \atop r-1} \right\}$, by taking one more step to the right as presented in Fig. 3.

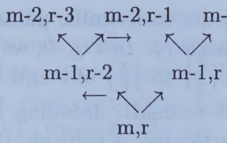


Figure 3a: Alternating decompositions

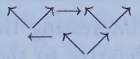


Figure 3b: Underlying structure

Note that in general alternating construction, we can no longer use RM codes. These only form the first "building blocks", such as $\left\{ {j \atop 0} \right\}$ and $\left\{ {j \atop j} \right\}$. By contrast, various nodes $\left\{ {j \atop i} \right\}$ only label the edges/paths that correspond to our new codes. The first simulation results obtained in this direction used an $(u, u+v)$-combination of $\left\{ {8 \atop 1} \right\}$ and $\left\{ {8 \atop 3} \right\}$ codes instead of the original $\left\{ {9 \atop 3} \right\}$ code. Even this simple combination improved the original code at low SNR. More sophisticated constructions similar to the one of Fig. 3a also outperform RM codes. However, the alternating constructions that we considered to date have not yet improved the performance of subcodes presented in Figures 4 to 9.

References

[1] Y.S. Han, C.R.P. Hartmann, and C.K. Mohan, "Efficient heuristic search algorithms for soft-decision decoding of linear block codes," *IEEE Trans. Inform. Theory,* vol. 44, pp. 3023-3038, 1998.

[2] G.A. Kabatyanskii, "On decoding of Reed-Muller codes in semicontinuous channels," *Proc. 2nd Int. Workshop "Algebr. and Combin. Coding Theory"*, Leningrad, USSR, 1990, pp. 87-91 (in Russian).

[3] R.E. Krichevskiy, "On the Number of Reed-Muller Code Correctable Errors," *Dokl. Soviet Acad. Sciences,* vol. 191, pp. 541-547, 1970.

[4] S.N. Litsyn, "Fast algorithms for decoding orthogonal and related codes," *Lecture Notes in Comp. Science,* no. 539, pp. 39-47, 1991.

[5] S.N. Litsyn, "On decoding complexity of low-rate Reed-Muller codes," *Proc. 9th All-Union Conf. on Coding Theory and Info. Transmission,* Part 1, Odessa, USSR, pp. 202-204, 1988 (in Russian).

[6] I.S. Reed, "A class of multiple error correcting codes and the decoding scheme," *IEEE Trans. Info. Theory,* vol. IT-4, pp.38-49, 1954.

[7] V. Sidel'nikov and A. Pershakov, "Decoding of Reed-Muller codes with a large number of errors," *Probl. Info. Transmission,* vol. 28, no. 3, pp. 80-94, 1992 (in Russian).

[8] G. Schnabl and M. Bossert, "Soft-decision decoding of Reed-Muller codes as generalized multiple concatenated codes," *IEEE Trans. Info. Theory,* vol. 41, pp. 304-308, 1995.

[9] I. Dumer, "Recursive decoding of Reed-Muller codes," *Proc. 37 Annual Allerton Conf. on Commun., Control, and Comp.,* Monticello, IL, Sept. 22-24, 1999, pp. 61-69.

FIG. 4: RM CODE (256, 37)

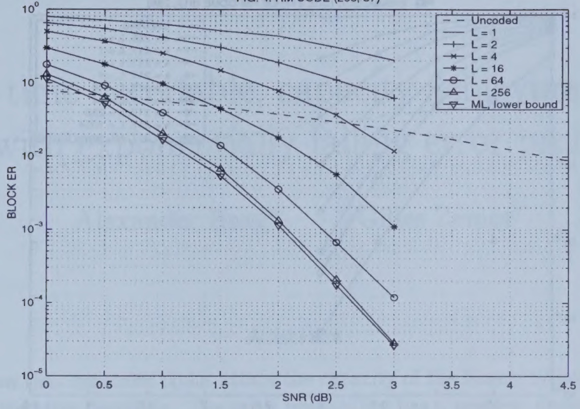


FIG. 5: SUBCODE (256, 78) OF RM CODE (256, 93)

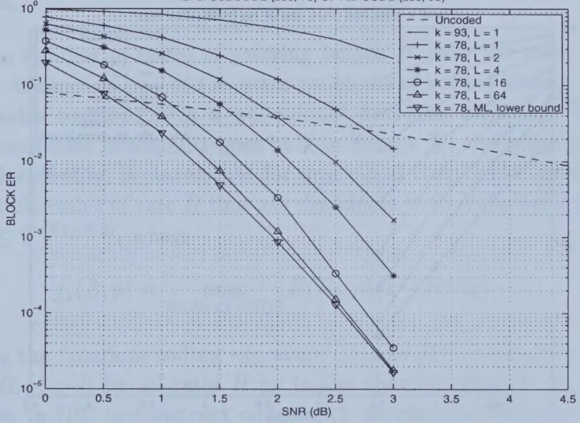


FIG. 6: SUBCODE (512, 101) OF RM CODE (512, 130)

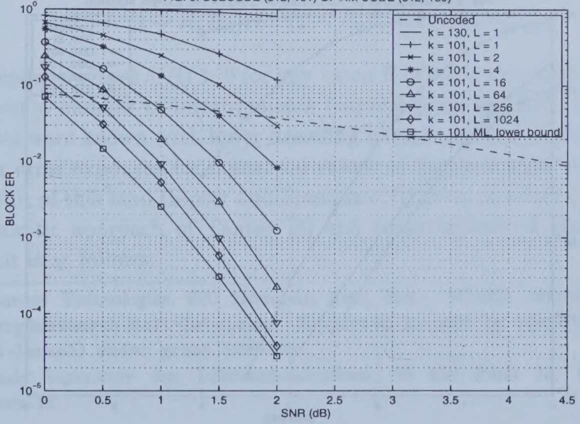


FIG. 7: SUBCODE (512, 217) OF RM CODE (512, 256)

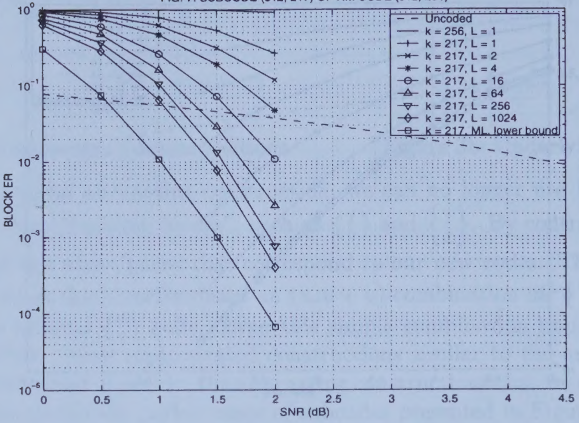


FIG. 8: SUBCODE (1024, 109) OF RM CODE (1024, 172)

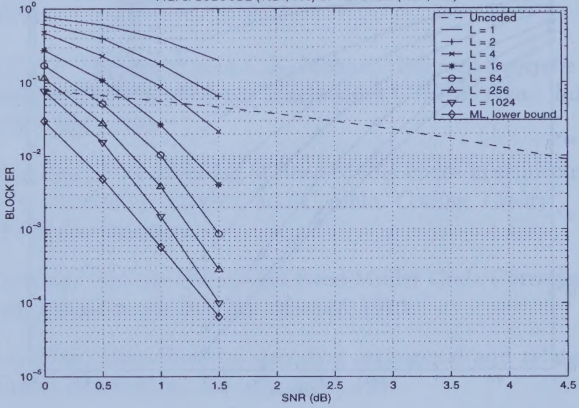
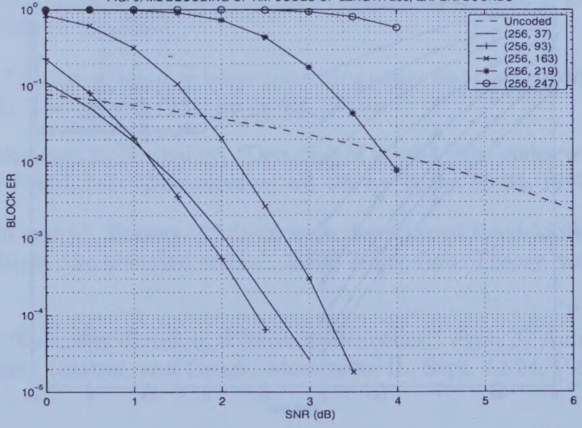


FIG. 9: ML DECODING OF RM CODES OF LENGTH 256, EXPER. BOUNDS



Linear-time decodable, capacity achieving binary codes with exponentially falling error probability

Alexander Barg* Gilles Zémor†

Abstract

We show that expander codes attain the capacity of the binary symmetric channel under iterative decoding. The error probability has a positive exponent for all rates between 0 and channel capacity. The decoding complexity grows linearly with code length.

Constructing families of codes of growing length N with low error probability and large minimum distance is one of the main problems of coding theory. The first result that gave us codes together with a polynomial decoding algorithm that achieves an exponentially small error probability goes back to Forney [3], and relies on a concatenated construction. Specifically, it states that for any $\varepsilon > 0$ there exists an infinite family of easily constructible codes of rate R that are decodable with complexity $O(N^2)$ and error probability $P_e \le 2^{-Nf_1(R,p)}$, where

$$f_1(R,p) = \max_{R \le R_0 < 1-H(p)} E(R_0,p)(1 - R/R_0) - \varepsilon. \tag{1}$$

Here $E(R_0,p)$ is the "random coding exponent" [4] and $H(\cdot)$ is the binary entropy function. Thus $f_1(R,p) > 0$ for all rates R up to the channel capacity. A similar idea was used by Zyablov in [10] to construct codes with polynomial decoding complexity and relative distance

$$\delta(R) = (1 - R/R_0)H^{-1}(1 - R_0) \quad (R \le R_0 \le 1). \tag{2}$$

Thus for these codes we have $\delta(R) > 0$ for any value R of the code rate $0 \le R < 1$. These results underwent a number of improvements (surveyed, for instance, in [2]), but until recently no codes were known with lower decoding complexity and nonvanishing relative distance and/or error exponent for nontrivial values of the code rate.

The first result of this kind (Sipser and Spielman [7]) gives us a family of codes, based on a graph-theoretic approach of Tanner [8] and constructions of Ramanujan graphs [5, 6]. The result is as follows:

---

*Bell Labs, Lucent Technologies, 600 Mountain Ave., Rm. 2C-375, Murray Hill, NJ 07974, USA. Email : `abarg@research.bell-labs.com`. Supported in part by the Binational Science Foundation (USA-Israel) under grant 1999099.

†École Nationale Supérieure des Télécommunications, 75 634 Paris 13, France. Email : `zemor@infres.enst.fr`

Theorem 1 [7] *For any $\varepsilon > 0$ there exists a polynomial-time constructible family of codes with distance $\delta - \varepsilon$ and rate $1 - 2H(\sqrt{\delta})$ for which any $\alpha < \delta/48$ fraction of errors can be corrected by a circuit of size $O(N \log N)$ and depth $O(\log N)$. The complexity of a sequential implementation of this decoding is $O(N)$.*

This result was a remarkable novelty because not only does it decrease the complexity of decoding but it was the first time that concatenation was not used to construct families of binary codes with a non-trivial rate and a non-zero relative minimal distance. It implies the existence of linear-time decodable codes with $R > 0$ for $0 \leq \delta < 0.0121$.

Sipser-Spielman's decoding algorithm was modified in [9] where the factor 48 was improved to 4. In the present paper we study the error probability for iterative decoding of expander codes and modified expander codes that we shall introduce. we shall use the decoding algorithm of [9].

Our main result is:

Theorem 2 *For a given rate R, any $\varepsilon > 0$ and $\alpha < 1$ there exists a polynomial-time constructible family of codes of length N such that $P_e(C,p) \leq 2^{-\alpha N f_2(R,p)}$, where*

$$f_2(R,p) = \max_{R \leq R_0 < 1-H(p)} E(R_0,p)(R_0 - R)/2 - \varepsilon.$$

The decoding complexity of these codes is the same as in Theorem 1.

Thus $f_2(R,p) > 0$ for all $0 \leq p < 1 - H(p)$. Comparing this result with (1), we see that $f_1(R,p) > f_2(R,p)$ for all R, p, so in terms of error probability expander codes do not surpass concatenated codes, but they do give an alternative to classical concatenation and they have a lower decoding complexity.

References

[1] N. Alon and F. R. K. Chung, "Explicit construction of linear sized tolerant networks," *Discrete Math.*, vol. 72, pp. 15–19, 1988.

[2] I. Dumer, "Concatenated codes and their multilevel generalizations," in *Handbook of Coding Theory* (V. Pless and W.C. Huffman, Eds.), Vol.II, Amsterdam: Elsevier Science (1998), pp. 1911-1988.

[3] G. D. Forney, Jr., *Concatenated Codes*, Cambridge, MA: MIT Press (1966).

[4] R. G. Gallager, *Information Theory and Reliable Communication*, New York: John Wiley & Sons (1968).

[5] A. Lubotsky, R. Philips and P. Sarnak, "Ramanujan graphs," *Combinatorica*, vol. 8, no 3, pp. 261–277, 1988.

[6] G. A. Margulis, "Explicit group theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators," *Problems Inform. Transm.*, vol. 24, no. 1, pp. 39–46, 1988.

[7] M. Sipser and D. A. Spielman, "Expander Codes", *IEEE Trans. on Inf. Theory*, IT-42, no 6, pp. 1710–1722, 1996.

[8] M. Tanner, "A recursive approach to low-complexity codes," *IEEE Trans. Inform. Theory*, IT-27, no 5, pp.533–547, 1981.

[9] G. Zémor, "On expander codes" to appear in *IEEE Trans. Inform. Theory*, Special issue on codes on graphs and iterative algorithms.

[10] V. V. Zyablov, "An estimate of complexity of constructing binary linear cascade codes," *Problems Inform. Transm.*, **7** (1) (1971), 3-10.

A Fundamental Need for Differentiated "Quality Of Service" Over Communication Links: An Information Theoretic Approach

Anant Sahai, Sanjoy Mitter *
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139
{sahai,mitter}@lids.mit.edu

October 3, 2000

Abstract

We give a simple example of a vector estimation problem over a specific noisy communication channel with feedback. Assuming a layered architecture in which the reliable transmission of bits is separated from the encoding of the source into bits, we apply information theoretic bounds to show that it is impossible to estimate this source in a mean-squared sense if all the bits are treated alike by the reliable transmission layer. This is true no matter how large a finite end to end delay we permit or how we encode the source. However, if the reliable transmission layer "splits" the channel into two appropriate streams with differentiated "Quality of Service," it becomes possible to achieve finite end-to-end mean-squared error by sending appropriate bits over the two bit pipes. It is argued that this fundamental need for differentiated service is more general than this particular example.

1 Introduction

"A bit is a bit is a bit" seems to have become the motto of this digital age. Though we often take it for granted, the existence of such a common currency for information is far from obvious. While in practice it has been justified by the rapid advancement of digital electronics and the rise of a communications infrastructure to match, its meaningfulness rests upon the fundamental theorems of information theory.

Broadly speaking, there are two sides to information theory. The channel capacity theorems establish the possibility for reliable transmission of bits using only unreliable media. On the other side, the rate distortion theorems establish fundamental tradeoffs between fidelity to a source and the average length of the encoding in bits. These two sides of information theory are joined together by the information transmission theorems relating to source/channel separation.

---

*This research supported by U.S. Army Grant PAAL03-92-G-0115, Center for Intelligent Control Systems.

These information transmission theorems provide the philosophical justification for both a layered architecture and the practice of treating all bits alike when it comes to reliable transmission. However, in many real world applications, especially "multimedia" ones, it has been found that it is sometimes conceptually useful to completely break the abstraction layer and do joint source/channel coding. Another angle, explored in Tse's Thesis [12], is to consider sources which vary more slowly than the acceptable delay and to use an adaptive quantization approach which keeps end-to-end delay small while still giving good distortion-rate performance.

The flip idea of "loosely coupled joint source/channel coding" [4] suggests that the reliable transmission layer should treat bit streams differently based on their meanings. Communication links are then evaluated not just by their bit-rates, but by other "Quality of Service" parameters as well. Even across a single link, it has been found useful to provide differentiated service for bit streams being sent at the same time. [16, 15] In all these examples, this discrepancy between theory and practice is explained by an appeal to economics or the technical need for finite small end-to-end delay in applications. While generally information theory only applies in the limit of large delays, new results suggest that there is more to the need for QoS than just the issue of small delays.

We have recently found a new information transmission theorem for scalar unstable Markov processes in both the control [7] and estimation [9] contexts. This theorem shows that transmitting scalar unstable Markov processes needs more from a communication link than just a given bit-rate C. It requires that the α parameter (a close relative to the error exponent) at that bit rate also be high enough. Insofar as it is used alongside the rate to evaluate a channel, the parameter α in the α-any-time capacity thus has the appearance of a fundamental "Quality of Service" term.

In this paper, we justify the QoS interpretation by considering a unstable vector source and the problem of estimating it in a mean-squared sense across a specific binary erasure channel with feedback. After reviewing some definitions in section 2, we introduce the basic problem setup in section 3, and show that it is impossible to reliably transmit the source over the channel in a mean-squared sense if we insist on treating all bits alike in the reliable transmission layer. In section 4, we show that providing differentiated service (with different α parameters for different bit streams) can allow us to achieve finite end-to-end mean squared error. This example shows that we do not need to specify an end-to-end delay constraint or any other user-level QoS requirement *a priori*. The need for QoS requirements other than rate and hence differentiated service can emerge from the nature of the source and the distortion measure.

2 Review

2.1 Notions of Capacity

Definition 2.1 *The* Shannon classical capacity C *of a channel is the least upper bound of the rates at which the channel can be used to transmit data with an arbitrarily small probability of error. [3]*

$$C = \sup\{R | \forall \epsilon > 0 \exists N, \mathcal{E}_N, \mathcal{D}_N P_{error}(\mathcal{E}_N, \mathcal{D}_N) < \epsilon, Rate(\mathcal{E}_N, \mathcal{D}_N) = R\}$$

The encoder/decoder pair $\mathcal{E}_N, \mathcal{D}_N$ has end-to-end delay less than or equal to N.

The above is the operational meaning of Shannon capacity and should be distinguished from the many ways of calculating what it is for various channels. [14] The implicit meaning of "reliable transmission" in Shannon capacity is the ability to get a small probability of error on each of the bits being sent. Recently, we proposed a stronger sense of reliable transmission and its induced notion of capacity for a channel: [9]

Definition 2.2 *The* α-any-time capacity $C_{at}(\alpha)$ *of a channel is the least upper bound of the rates at which the channel can be used to transmit data with an arbitrarily small probability of error that decays with delay at least exponentially at a rate* α.

$$C_{at}(\alpha) = \sup\{R|\exists(K > 0, Rate(\mathcal{E},\mathcal{D}^a) = R)\,\forall d > 0\ P_{error}(\mathcal{E},\mathcal{D}^a, Delay = d) \leq K2^{-\alpha d}\}$$

This operational definition does not tell how to calculate α-any-time capacity. It is closely related to the traditional ideas of error exponents [3] but imposes a stronger restriction since it requires the encoder $\mathcal{E}$ to be the same for all end-to-end delays. This requirement arises because the bits to be transmitted are being generated in real-time as opposed to our transmitting a long message known in advance.

Viewing this from the decoder's perspective tells us why we are calling it any-time capacity. We can interrupt the decoding process for a given bit at "any time" and the resulting answer is increasingly meaningful the longer we wait. The α specifies the rate at which we want the answers to improve. The result is:

Lemma 2.1 *[8] Almost every bit is eventually decoded correctly. Let* $\tilde{S}_j^d = \hat{S}_j$ *when using delay d. For any* $\alpha > 0$, *if* $R < C_{at}(\alpha)$ *then* $\exists(\mathcal{E},\mathcal{D}^a)$ *such that for all* $j > 0$ *the sequence* $[\tilde{S}_j^0, \tilde{S}_j^1, \ldots]$ *eventually converges to the correct value* S_j *and stays there with probability one.*

This property of getting every bit correct is analogous to the way in which TCP/IP provides reliable transmission over the Internet. [6]

2.2 Sources and Information Transmission

A Markov source $\{X_t\}$ with dynamics A and driven by noise $\{W_t\}$ is defined by the following:

$$X_{t+1} = AX_t + W_t \tag{1}$$

where A, X_t, and W_t are all of the appropriate dimensions for the recurrence to make sense. In the scalar case, they are all one dimensional.

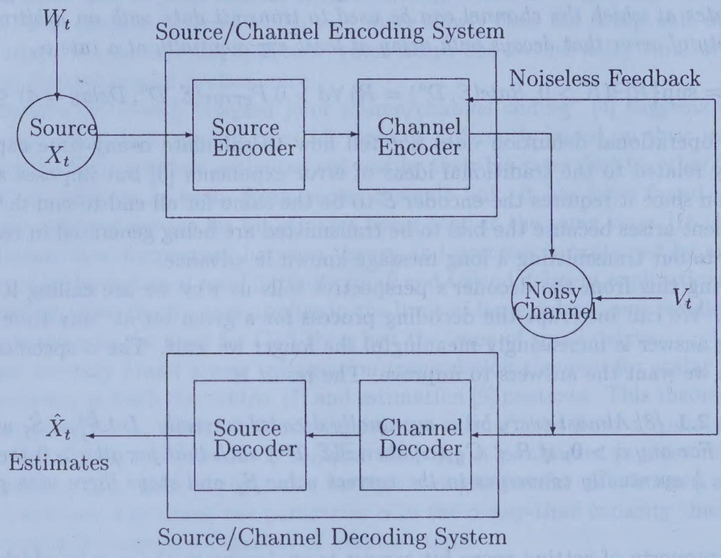
The natural problem of transmitting such a source across a noisy channel is the following: *Is it possible to design encoders and decoders within any specified end-to-end delay constraint so that the output of the decoder* $\{\hat{X}_t\}$ *achieves a desired mean-squared performance* $\sup_t E[\|\hat{X}_t - X_t\|^2] = K$*?* For the cases where A is a stable matrix (with all eigenvalues having magnitude less than 1), the problem has a well established complete solution [1] which can be naturally extended to a wide range of channels using the general source/channel separation results in [13].

In the case where the process is unstable (A having an eigenvalue with magnitude greater than 1), the first question is whether it is possible to get a finite K at all. After all, such unstable processes have asymptotically infinite variance. For the unstable scalar case with bounded noise, [9] provides the answer: (Proofs in [8])

Theorem 2.3 *If the source is scalar and driven by bounded noise, the source in (1) can be tracked with finite mean squared error across a noisy channel iff there is an* $\epsilon > 0$ *for which* $C_{at}(2\log_2 A + \epsilon) > \log_2 A$ *for the channel.*

3 A Simple Vector Problem

Consider a particular vector source and the problem of transmitting it with finite mean squared error across a binary erasure channel with noiseless feedback.


W_t
Source/Channel Encoding System
Noiseless Feedback
Source
X_t
Source Encoder
Channel Encoder
Noisy Channel
V_t
$\hat{X}_t$
Estimates
Source Decoder
Channel Decoder
Source/Channel Decoding System

The specific vector source we consider has $X_t \in \Re^5$, a bound on the driving noise $\|W_t\|_\infty \leq \frac{\Omega}{2}$, and known initial condition $X_0 = 0$. It is defined by the matrix of dynamics

$$A = \begin{pmatrix} 1.178 & 0 & 0.04 & 0 & 0.04 \\ 1.08 & 1.058 & 0.36 & 0 & 0.36 \\ 0.36 & 0 & 1.178 & 0 & 0.12 \\ -0.12 & 0 & -0.04 & 1.058 & -0.04 \\ -0.12 & 0 & -0.04 & 0 & 1.018 \end{pmatrix} \quad (2)$$

In our simple example, we consider a binary erasure channel with erasure probability $e = 0.27$. The binary erasure channel models situations where errors can be reliably detected at the receiver. In the model, sometimes the single bit being sent does not make it through with probability e, but otherwise it makes it through correctly.

It can be shown that the Shannon classical capacity of this channel is $1 - e$ bits per channel use regardless of whether the encoder has feedback or not. [3] For the case where the encoder has access to instantaneous noiseless feedback, the anytime capacity can also be evaluated and is given by:

Theorem 3.1 *([7], Proof in [8]) For the binary erasure channel with noiseless feedback and probability of erasure e, let η range over $(0, \infty)$:*

$$C_{at}(\eta - \log_2(1 + (2^\eta - 1)e)) = 1 - \frac{1}{\eta}\log_2(1 + (2^\eta - 1)e)$$

3.1 Transform Coding

Looking at the source process in transformed coordinates is often of value in lossy coding. [5] In our case, it is illuminating to consider the transformed variable $\tilde{X}_t = TX_t$ where

$$T = \begin{pmatrix} 3 & 0 & 1 & 0 & 1 \\ 0 & 1 & -2 & 0 & 3 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In that case, the transformed dynamics are given by a new $\tilde{A} = TAT^{-1}$ matrix:

$$\tilde{A} = \begin{pmatrix} 1.258 & 0 & 0 & 0 & 0 \\ 0 & 1.058 & 0 & 0 & 0 \\ 0 & 0 & 1.058 & 0 & 0 \\ 0 & 0 & 0 & 1.058 & 0 \\ 0 & 0 & 0 & 0 & 1.058 \end{pmatrix} \tag{3}$$

The initial condition is unchanged $\tilde{X}_0 = 0$ and the new driving noise $\tilde{W}_t$ has a new bound given by $\|\tilde{W}_t\|_\infty \leq 3\Omega$. It is also clear that having an encoder/decoder pair which achieves a finite mean squared error between the original $\{X_t\}$ and $\{\hat{X}_t\}$ processes is equivalent to having a pair which achieves a finite mean squared error between the transformed $\{\tilde{X}_t\}$ process and its corresponding transformed reconstruction.

Furthermore, being able to achieve a finite mean squared error with any finite end-to-end delay implies that we are able to achieve a finite mean squared error with no delay. This is because we could use an estimate for X_{t-d} to give an estimate for X_t by simply premultiplying it by A^d. After all,

$$\begin{aligned} E\left[\|X_t - A^d\hat{X}_{t-d}\|^2\right] &= E\left[\|(A^d X_{t-d} + \sum_{i=1}^{d} A^{i-1}W_{t-i}) - A^d\hat{X}_{t-d}\|^2\right] \\ &\leq E\left[\|A^d(X_{t-d} - \hat{X}_{t-d})\|^2\right] + E\left[\|\sum_{i=1}^{d} A^{i-1}W_{t-i}\|^2\right] + \\ &\quad 2E\left[\|A^d(X_{t-d} - \hat{X}_{t-d})\|\|(\sum_{i=1}^{d} A^{i-1}W_{t-i}\|\right] \end{aligned}$$

For any finite d, the three terms on the right hand side are all bounded for all t if we can achieve finite mean squared error for delay d. Hence the sum is as well.

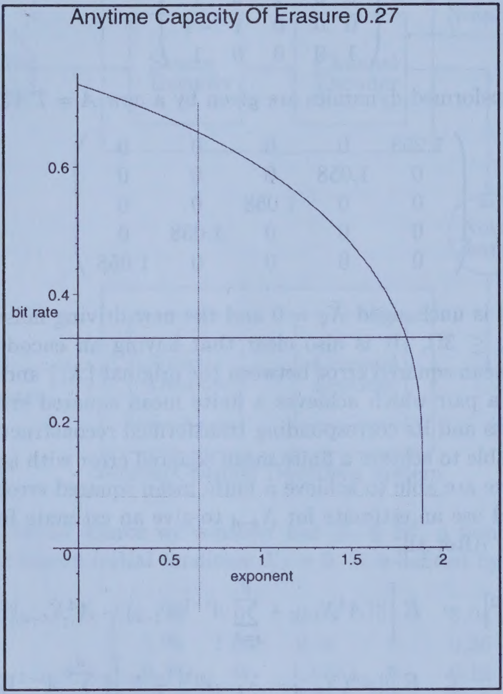
3.2 Fundamental Requirements

A necessary condition for achieving finite mean squared error for a vector valued process is getting a finite mean squared error for all the components. In the transformed domain, the dynamics of the components of our example are like those of five separate scalar unstable Markov sources. One of these is fast, and the four others are not as fast. Therefore, our Theorem 2.3 applies and we can easily conclude that we need the channel to satisfy both:

$$\begin{aligned} C_{\text{at}}(2\log_2(1.258)) &> \log_2(1.258) \\ C_{\text{at}}(2\log_2(1.058)) &> \log_2(1.058) \end{aligned}$$

Moreover, our work on sequential rate distortion [10, 11] tells us that the total rate must be larger than the the sum of the logs of the unstable eigenvalues, giving us the additional condition that:

$$C > \log_2(1.258) + 4\log_2(1.058)$$



It is easy to check that all these requirements are satisfied individually by the binary erasure channel with erasure probability $e = 0.27$. However as stated, these are only necessary conditions, not sufficient ones. In a sense, we need them all satisfied simultaneously for the different bitstreams in question.

3.3 Treating All Bits Alike

By using techniques from [8, 11], it is easy to construct recursive source codes for each of the five components of the transformed source. As long as we allow ourselves $R_1 > \log_2(1.258)$ bits per unit time for encoding the first component and $R_{2,3,4,5} > \log_2(1.1)$ bits per unit time time on the others, we can achieve finite mean squared error assuming no noise in transmitting these bits.

If we follow a strictly layered strategy and then require that all these bits be treated identically by the reliable transmission layer, we come up against a problem. For an erasure channel with $e = 0.27$, The maximum α for which the anytime capacity is larger than $(\log_2(1.258) + 4\log_2(1.058))$ is only around 0.646. This is enough for the four slow components which each require $\alpha > 2\log_2(1.058) \approx 0.163$. But it is less than the minimum α we require for the fast component: $\alpha > 2\log_2(1.258) \approx 0.662$. As long as we insist on treating all the bits alike in the reliable transmission layer, it is impossible to achieve a finite mean squared error on the first component of the transformed source. Because this is a consequence of our fundamental separation theorem for such processes, it is true regardless of how much end-to-end delay we are willing to tolerate or how we do the source encoding.

4 Differentiated Service

Consider an approach more in the spirit of "loosely coupled joint source/channel coding." [4] We get a higher priority for the bitstream representing the "faster" component requiring $\alpha > 2\log_2(1.258)$ by using the following strategy for reliable transmission:

- Store the incoming bits from the different streams into prioritized buffers — one buffer for each distinct priority level.
- At every opportunity for channel use, transmit the oldest bit from the highest priority buffer that is not empty.
- If the bit was received correctly, remove it from the appropriate buffer.

For our transformed source, we can use the scalar codes from [11, 8] to encode the first component at a rate $R_1 = \frac{1}{3} > \log_2(1.258)$ and assign this bitstream the higher priority. The first stream generates a bit every third time step. The other components are all encoded at a rate $R_{2,3,4,5} = \frac{1}{12} > \log_2(1.058)$ and gets the lower priority. These together also generates a bit every three steps (4 bits every 12 time steps). The total rate is therefore $R = \frac{2}{3} < 1 - e = 0.73$.

Since there is noiseless feedback and the incoming bitstreams are deterministic in their timing, the decoder can keep track of the encoder's buffer sizes and thus knows which incoming bit belongs to which stream. The source decoder takes all available bits and makes the best prediction of where it thinks the source process is.

4.1 The Source Codes

Our focus is not on optimizing performance and so we will just use a simple recursive source code [11, 8] which keeps the errors strictly bounded. An approach that asymptotically achieves optimal performance (rate-distortion bounds) can be found in [8]. Here, we simply apply the following theorem to each component of the transformed source.

Theorem 4.1 *[8] For all $\epsilon > 0$, every scalar discrete-time unstable linear Markov process with parameter $a > 1$ driven by bounded noise $-\frac{\Omega}{2} \le W_t \le \frac{\Omega}{2}$, can be tracked by an encoder F^R and decoder G^R with rate $R \le \log_2 a + \epsilon$. Moreover, there exists a constant ν depending only on (a, Ω, R) such that $|X_t - \hat{X}_t| \le \nu$ for all t.*

The source code works by encoding the predictive error signal $(X_{t+1} - a\hat{X}_t)$ and keeps it within a box of $[-\nu, \nu]$. Whenever the error signal threatens to leave that box, the encoder transmits a bit. Since each encoded bit cuts the uncertainty by a factor of 2, while the dynamics make it grow by a factor of a at each time, such a code can work for any $R > \log_2 a$. The closer R gets to that limit, the larger ν has to be. At the receiver's side, we use all the bits we have received to make our estimates. If we are missing some bits, we just predict by multiplying what we have by the appropriate power of that component's eigenvalue.

4.2 Analyzing The Channel Code

Notice that for any delay d time units, the channel decoder makes an error for bitstream i only if the encoder's buffer i contains more than dR_i bits still awaiting transmission. If there are less bits waiting, it means that the bit from d time units ago has already made it across. Thus, to analyze how the average probability of error varies with delay, we only need to study the steady-state distribution of the number of bits in each buffer.

4.2.1 The High Priority Stream

Since the highest priority stream preempts all lower priority streams, it effectively does not have to share the channel at all. So we can study its queue length using a simple Markov chain by blocking time into three time unit blocks. Then, the number of bits awaiting transmission at the end of a block is the Markov state with transition probabilities $p_{i,j}$ representing the probability that state i will go next to state j.

$$\begin{aligned}
p_{0,0} &= 3e^2(1-e) + 3e(1-e)^2 + (1-e)^3 \\
p_{i,i+1} &= e^3 \\
p_{i,i} &= 3e^2(1-e) \\
p_{i,i-1} &= \begin{cases} 3e(1-e)^2 + (1-e)^3 & \text{if } i = 1 \\ 3e(1-e)^2 & \text{if } i > 1 \end{cases} \\
p_{i,i-2} &= (1-e)^3
\end{aligned}$$

It is possible to calculate the steady state distribution π for this Markov chain. By some algebraic manipulation we can get the following recurrence relation:

$$(1-e)^3\pi_i = \begin{cases} (1-3e^2(1-e))\pi_{i-2} - 3e(1-e)^2\pi_{i-1} - e^3\pi_{i-3} & \text{if } i > 2 \\ (1-3e^2(1-e) - 3e(1-e)^2 - 3e^2(1-e))\pi_{i-2} - 3e(1-e)^2\pi_{i-1} & \text{if } i = 2 \end{cases}$$

It turns out that $\pi_i \propto \left(\frac{2e^3}{1+2e^3+(1-e)\sqrt{1+2e-3e^2}-3e^2}\right)^i$ as i gets large and thus:

$$\begin{aligned}
P_{\text{error}}(\text{Delay} = d) &\leq P(\text{Buffer State} > dR_1) \\
&\leq K\left(\frac{2e^3}{1+2e^3+(1-e)\sqrt{1+2e-3e^2}-3e^2}\right)^{\frac{d}{3}} \\
&= K2^{-\frac{1}{3}\log_2\left(\frac{1+2e^3+(1-e)\sqrt{1+2e-3e^2}-3e^2}{2e^3}\right)d}
\end{aligned}$$

which for $e = 0.27$ results in a fast enough $\alpha \approx 1.799 > 2\log_2(1.258) = 0.662$.

4.2.2 The Low Priority Streams

Notice that regardless of the realization of channel noise, the sum of the queue lengths for the two buffers is identical to the queue length for a hypothetical single stream at the combined rate. So it suffices to look at a single stream with rate $\frac{2}{3}$. Once again, we will group channel uses into threes so that the number of bits awaiting transmission at the end of a group is the Markov state for the system. This gives us:

$$\begin{aligned}
p_{0,0} &= 3e(1-e)^2 + (1-e)^3 \\
p_{i,i+2} &= e^3
\end{aligned}$$

$$\begin{aligned} p_{i,i+1} &= 3e^2(1-e) \\ p_{i,i} &= 3e(1-e)^2 \\ p_{i,i-1} &= (1-e)^3 \end{aligned}$$

The steady state distribution π for the state in this case can be calculated just as before. By some algebraic manipulation we can get the following similar recurrence relation for it:

$$(1-e)^3\pi_i = \begin{cases} (1-3e(1-e)^2)\pi_{i-1} - 3e^2(1-e)\pi_{i-2} - e^3\pi_{i-3} & \text{if } i > 2 \\ (1-3e(1-e)^2)\pi_{i-1} - 3e^2(1-e)\pi_{i-2} & \text{if } i = 2 \\ (1-3e(1-e)^2 - 3e^2(1-e))\pi_{i-1} & \text{if } i = 1 \end{cases}$$

For large, i, we then have $\pi_i \propto \left(\frac{2e^2}{2e^2+\sqrt{4e-3e^2}-3e}\right)^i$. This rate of decay is clearly much slower than the one for the higher priority queue and thus, for large queue lengths, it dominates. Therefore we have for the bits in the lower priority queue:

$$\begin{aligned} P_{\text{error}}(\text{Delay} = d) &\leq P(\text{Combined Buffer State} > d(R_2+R_3+R_4+R_5)) \\ &\leq K\left(\frac{2e^2}{2e^2+\sqrt{4e-3e^2}-3e}\right)^{\frac{d}{3}} \\ &= K2^{-\frac{1}{3}\log_2(\frac{2e^2+\sqrt{4e-3e^2}-3e}{2e^2})d} \end{aligned}$$

which for $e = 0.27$ results in an $\alpha \approx 0.285 > 2\log_2(1.058) \approx 0.163$ so our scheme is fast enough for all the slow components as well!

The direct part of Theorem 2.3 then shows that the cascaded source and channel codes achieve a finite mean squared error for all components of the transformed source. Thus, the system achieves a finite mean squared error on the original source as well.

5 Conclusion and Speculations

We have presented a simple example of a vector source and a specific channel for which it is impossible to achieve finite end-to-end mean-squared-error without using some form of differentiated service at the reliable transmission layer. Furthermore, in order to evaluate the bit-pipes provided by the reliable transmission layer, we have shown that the ideas of α-anytime-capacity are relevant and have used our previous separation results [9] to motivate a scheme which does achieve finite end-to-end performance.

It is hoped that this work goes some small distance towards meeting the challenge layed down by Ephremides and Hajek: [2]

> The interaction of source coding with network-induced delay cuts across the classical network layers and has to be better understood. The interplay between the distortion of the source output and the delay distortion introduced on the queue that this source output feeds into may hold the secret of a deeper connection between information theory.

This toy example with a single simple link opens the door towards a real Information Theory of "Quality of Service." Even though in the real world, few processes of interest actually have asymptotically infinite variance, we conjecture that these ideas will be a useful approximation whenever the upper limit of the tolerable end-to-end delay is within the range of time for which an unstable model is applicable. This may be a useful abstraction that underlies many "real-time" data sources. We suspect that this is true for not just control problems, but many others (including multimedia ones) as well.

References

[1] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[2] A. Ephremides and B. Hajek, "Information Theory and Communication Networks: An Unconsummated Union." IEEE Trans. on Information Theory, 44(6):2416–2434, 1998.

[3] R. Gallager, *Information Theory and Reliable Communication.* New York, NY: John Wiley and Sons, 1971.

[4] P. Haskell, D. G. Messerschmitt, and L. Yun, "Architectural Principles for Multimedia Networks," in *Wireless Communications: Signal Processing Perspectives*, H. V. Poor and G. W. Wornell, editors, Upper Saddle River, NJ: Prentice-Hall, 1998.

[5] A. K. Jain, *Fundamentals of Digital Image Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.

[6] J. Postel, "Transmission Control Protocol", RFC 793, USC/Information Sciences Institute, September 1981 (http://www.ietf.org/rfc/rfc0793.txt)

[7] A. Sahai, "Evaluating Channels for Control: Capacity Reconsidered." Proc. 2000 American Control Conference, 2000.

[8] A. Sahai, "Any-time Information Theory." PhD Dissertation in progress.

[9] A. Sahai, ""Any-time" Capacity and A Separation Theorem For Tracking Unstable Processes," Proc. 2000 IEEE International Symposium on Information Theory, 2000.

[10] S. Tatikonda, A. Sahai, S. Mitter, "Control of LQG Systems Under Communication Constraints." Proc. 37th IEEE Conference on Decision and Control, 1998.

[11] S. Tatikonda, "Control Under Communication Constraints." PhD Dissertation, Massachusetts Institute of Technology, 2000.

[12] D. N. C. Tse, "Variable-rate Lossy Compression and its Effects on Communication Networks," PhD Dissertation, Massachusetts Institute of Technology, 1995.

[13] S. Vembu, S. Verdu, and Y. Steinberg, "The Source-Channel Separation Theorem Revisited." IEEE Trans. on Information Theory, 41(1):44–54, 1995.

[14] S. Verdu and T. S. Han, "A General Formula for Channel Capacity." IEEE Trans. on Information Theory, 40(4):1147–1157, 1994.

[15] X. Xiao and L. M. Ni, "Internet QoS: a big picture," IEEE Network, 13(2):8–18, 1999.

[16] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: a new resource ReSerVation Protocol," IEEE Network, 7(5):8–18, 1993.

Channel Coding with Feedback

Sekhar Tatikonda and Sanjoy Mitter
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

In this paper we provide a general coding theorem for channels with feedback. To this end we introduce Massey's notion of directed mutual information. [Mas] We next show that the problem of feedback coding for Markov channels can be cast as a partially observed stochastic control problem. Consequently we can use the tools of dynamic programming to solve the directed mutual information optimization problem underlying the capacity problem.

1 Introduction

Because of the increased demand for wireless communication and networked systems there is a renewed interest in the problem of feedback channel coding. Feedback can both increase the capacity of a noisy channel and decrease the complexity of the encoder and decoder.

Recently Verdu and Han gave a very general formulation of the channel coding problem without feedback. [VH] Here we generalize that formulation to the case of channels with feedback. Along the way we introduce the notion of directed mutual information and argue that it provides the correct measure of capacity. Directed mutual information was introduced by Massey who attributes it to Marko. [Mas], [Mar]

The Verdu-Han result is quite general. But that generality comes at a cost. It can be very difficult to solve the capacity optimization problem. One of the difficulties is due to a lack of information stability. We examine the class of Markov channels. We show that the problem of feedback coding for Markov channels can be cast as a partially observed stochastic control problem. Consequently we can use the tools of dynamic programming to solve the directed mutual information optimization problem underlying the capacity problem. We give a clean formulation for computing the capacity for a large class of channels with memory and different forms of feedback.

In section 2 we present the formulation and introduce directed mutual information. We prove a channel coding theorem with feedback in section 3. Section 4 is devoted to channels with state. Because of space restrictions not all proofs will be carried out in full. See [Tat] for full details.

2 Channel Capacity and Feedback

Here we define channels, encoders, and decoders. Furthermore we define the operational channel capacity, introduce directed information, and define feedback capacity.

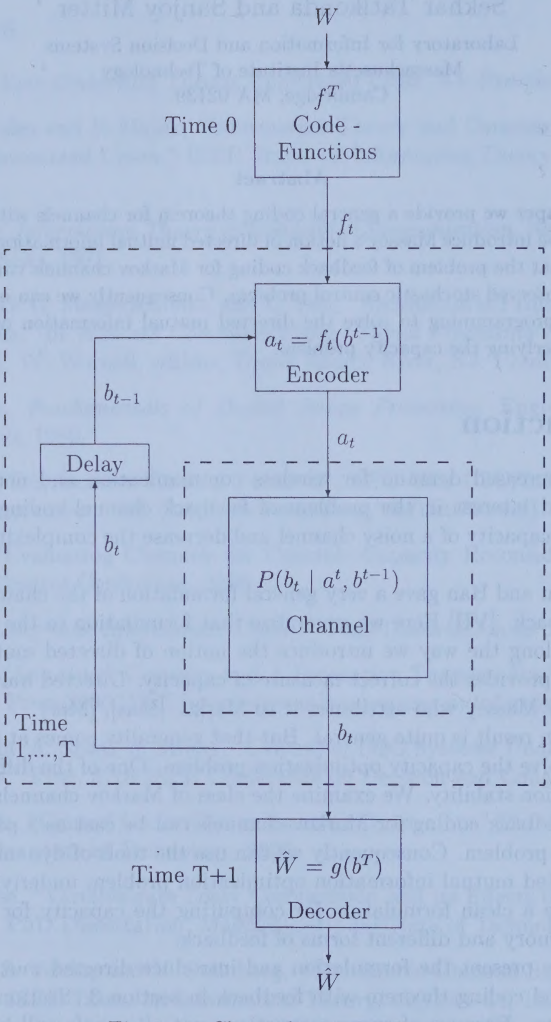

W
Time 0
f^T
Code -
Functions
f_t
a_t = f_t(b^{t-1})
Encoder
b_{t-1}
a_t
Delay
b_t
P(b_t | a^t, b^{t-1})
Channel
Time
1,...,T
b_t
Time T+1
Ŵ = g(b^T)
Decoder
Ŵ

Figure 1: Channel with Feedback

Let $\mathcal{A}, \mathcal{B}$ be spaces with a finite number of elements. Let A_t, B_t be measurable random variables taking values in $\mathcal{A}, \mathcal{B}$ respectively. The product spaces $\mathcal{A}^T$ and $\mathcal{B}^T$ represent the input and output spaces for the channel for a time horizon T. See figure 1.

Given a sequence of variable $x_1, ..., x_T$ we use the notation $x^T \triangleq (x_1, ..., x_T)$.

Definition 2.1 *A channel is a family of stochastic kernels* $\{P_{B_t|A^t,B^{t-1}}\}_{t=1}^T$. *(T may be infinity.)*

Definition 2.2 *A message set is a set* $\mathcal{W} = \{1, ..., M\}$.

Definition 2.3 *A channel code-function is a sequence of T deterministic measurable maps* $\{f_t\}_{t=1}^T$ *such that* $f_t : \mathcal{B}^{t-1} \to \mathcal{A}$ *which takes* $b^{t-1} \mapsto a_t$. *Denote the set of all code-functions by* $\mathcal{F}_T = \{f^T\}$.

Definition 2.4 *A channel encoder, or channel code, is a set of M channel code-functions. Denote them by* $f^T[w] = \{f_1[w], ..., f_T[w]\}, \quad w = 1, ..., M$.

Definition 2.5 *A channel input distribution is a sequence of stochastic kernels* $\{P_{A_t|A^{t-1},B^{t-1}}\}_{t=1}^T$.

Let $P_{\mathcal{F}^T}$ be a distribution on $\mathcal{F}_T$. We now show that $P_{\mathcal{F}^T}$ and the channel $\{P_{B_t|A^t,B^{t-1}}\}_{t=1}^T$ uniquely define a joint measure $P_{\mathcal{F}^T,A^T,B^T}$ on $\mathcal{F}_T \times \mathcal{A}^T \times \mathcal{B}^T$.

Lemma 2.1 *$P_{\mathcal{F}^T}$ and the channel $\{P_{B_t|A^t,B^{t-1}}\}_{t=1}^T$ uniquely define a measure $P_{\mathcal{F}^T,A^T,B^T}$ on $\mathcal{F}_T \times \mathcal{A}^T \times \mathcal{B}^T$.*

Proof: Assume (f^T, a^T, b^T) satisfy $a_t = f_t(b^{t-1})\ \forall t$. (If it did not then the probability would be zero.) Then

$$\begin{aligned} P(f^T, a^T, b^T) &= P(f^T) \prod_{t=1}^T P(b_t \mid a^t, b^{t-1}, f^T) P(a_t \mid a^{t-1}, b^{t-1}, f^T) \\ &= P(f^T) \prod_{t=1}^T P(b_t \mid a^t, b^{t-1}) \end{aligned}$$

□

Definition 2.6 *A channel decoder is a map* $g : \mathcal{B}^T \to \mathcal{W}$ *taking* $b^T \mapsto w$.

2.1 Operational Channel Capacity

We take the distribution, P_W, on $\mathcal{W}$ to be the uniform distribution.

Definition 2.7 *A (T, M, ϵ) channel code over time horizon T has a channel code with M code-functions, a decoder g, and an error probability* $\frac{1}{M}\sum_{w=1}^M \Pr(w \neq g(b^T)|w) \leq \epsilon$.

The superscript "o" represents the word "operational."

Definition 2.8 *R is an ϵ-achievable rate if, for every $\delta > 0$ there exists, for sufficiently large T, an (T, M, ϵ) channel code with rate $\frac{\log M}{T} > R - \delta$. The maximum $\epsilon -$ acheivable rate is the called the $\epsilon-$capacity and denoted C_ϵ^o. The channel capacity is defined as the maximal rate that is achievable for all $0 < \epsilon < 1$ and is denoted C^o.*

In the case when there is no feedback the traditional next step is to introduce a particular mutual information optimization. The value of this optimization turns out to equal the operational capacity. This no longer holds for the case with feedback.

2.2 Directed Mutual Information and Capacity

It turns out that mutual information is insufficient for computing the channel capacity of feedback channels. [Mas] To appreciate this fact we need to distinguish between causal independence and probabilistic independence. Under feedback it is no longer true that $P_{B_t|A^T,B^{t-1}} = P_{B_t|A^t,B^{t-1}}$. Now $I(A^T;B^T)$ depends on $P_{B_t|A^T,B^{t-1}}$ whereas we would like a measure that is only dependent on $P_{B_t|A^t,B^{t-1}}$. That measure turns out to be the directed mutual information. [Mas]

Definition 2.9 $I(A^T \to B^T) \triangleq \sum_{t=1}^T I(A^t; B_t|B^{t-1})$

Note that on the right-hand side the superscript "t" in the A term. For the regular mutual information there would be a superscript "T".

Definition 2.10 *Define* $\vec{P}(b^T|a^T) \triangleq \prod_{t=1}^T P(b_t|a^t, b^{t-1})$ *and* $\vec{P}(a^T|b^T) \triangleq \prod_{t=1}^T P(a_t|a^{t-1}, b^{t-1})$

One can show that $I(A^T \to B^T) = D(P_{A^T,B^T} \,\|\, \vec{P}_{A^T|B^T} P_{B^T})$. [Tat]

Lemma 2.2 *If we are using a channel encoder without feedback or a channel input distribution without feedback then* $I(A^T; B^T) = I(A^T \to B^T)$.

Proof: See theorem 1 of [Mas]. □

The following is the main technical lemma we will need to prove the coding theorem.

Lemma 2.3 *We are given a joint measure* P_{F^T,A^T,B^T} *such that* $a_t = f_t(b^{t-1})$. *Then the following two ratios are equal*

$$\frac{P_{F^T,B^T}(f^T,b^T)}{P_{F^T}P_{B^T}(f^T,b^T)} = \frac{P_{A^T,B^T}(a^T,b^T)}{\vec{P}_{A^T|B^T}P_{B^T}(a^T,b^T)}$$

Proof: First note that $P_{B_t|B^{t-1},F^T}(b_t|b^{t-1}, f^T) = P_{B_t|B^{t-1},A^t}(b_t|b^{t-1}, a^t)$ where $a_t = f_t(b^{t-1})$. Then

$$\begin{aligned}
\frac{P_{F^T,B^T}(f^T,b^T)}{P_{F^T}P_{B^T}(f^T,b^T)} &= \frac{\prod_{t=1}^T P_{B_t|B^{t-1},F^T}(b_t|b^{t-1},f^T)}{P_{B^T}(b^T)} \\
&= \frac{\prod_{t=1}^T P_{B_t|B^{t-1},A^t}(b_t|b^{t-1},a^t)}{P_{B^T}(b^T)} \\
&= \frac{\vec{P}_{B^T|A^T}(b^T|a^T)\vec{P}_{A^T|B^T}(a^T|b^T)}{P_{B^T}(b^T)\vec{P}_{A^T|B^T}(a^T|b^T)} \\
&= \frac{P_{A^T,B^T}(a^T,b^T)}{\vec{P}_{A^T|B^T}P_{B^T}(a^T,b^T)}
\end{aligned}$$

□

2.2.1 Feedback Channel Capacity

When computing the capacity of a channel it will turn out that we are most interested in the convergence properties of the random variables $\frac{1}{T}\log\frac{P_{A^T,B^T}(A^T,B^T)}{\vec{P}_{A^T|B^T}P_{B^T}(A^T,B^T)}$. If there are reasonable regularity properties, like the ergodicity of the underlying process, then these random variables will converge in probability to a deterministic limit. In the absence of any such structure we are forced to follow Verdu and Han's lead and define the following "floor" limit. [VH] The *liminf in probability* of a sequence of random variables $\{X_t\}$ is defined as the largest extended real number α such that $\forall \epsilon > 0$ $\lim_{t\to\infty} \Pr[X_t \leq \alpha - \epsilon] = 0$. Denote this number α as $\underline{X}$.

Definition 2.11 *For a sequence of joint measures* $\{P_{A^T,B^T}\}_{T=1}^{\infty}$ *let* $\underline{I}(A \to B) \triangleq \liminf_{in\ prob} \frac{1}{T} \log \frac{P_{A^T,B^T}(A^T,B^T)}{\vec{P}_{A^T|B^T} P_{B^T}(A^T,B^T)}$.

Definition 2.12 *Let* $\mathcal{S}_T = \{\{P_{A_t|A^{t-1},B^{t-1}}\}_{t=1}^{T}\}$ *be the set of all channel input distributions.*

Definition 2.13 *For finite* T *let* $C_T = \sup_{S \in \mathcal{S}_T} \frac{1}{T} I(A^T \to B^T)$. *For the infinite horizon let* $C = \sup_{S \in \mathcal{S}_\infty} \underline{I}(A \to B)$.

We now show that C is in fact the operational channel capacity with feedback. Specifically $C^{\mathrm{O}} = C$.

3 Feedback Channel Coding Theorem

In this section we prove the channel coding theorem. We proceed first with the converse part and then proceed to the direct part. When there is no feedback we recover the results in [VH].

3.1 Converse Theorem

Choose a (T, M, ϵ) channel code $\{f^T(w)_{w=1}^M\}$. Place a prior probability $\frac{1}{M}$ on each code-function $f^T(w)$. By lemma 2.1 this defines a joint measure P_{W,A^T,B^T}. The following is a generalization of the Verdu-Han converse presented in [VH].

Lemma 3.1 *Every* (T, M, ϵ) *channel code satisfies*

$$\epsilon \geq P_{A^T,B^T}\left(\frac{1}{T} \log \frac{P_{A^T,B^T}(a^T,b^T)}{\vec{P}_{A^T|B^T} P_{B^T}(a^T,b^T)} \leq \frac{1}{T} \log M - \gamma\right) - e^{-\gamma T} \qquad \forall \gamma > 0$$

Proof: Choose a $\gamma > 0$. Let $D_w \subset \mathcal{B}^T$ be the decoding region for message w. The only restriction we place on the decoding regions is that they do not intersect: $D_w \cap D_{\tilde{w}} = \emptyset \ \forall \tilde{w} \neq w$. (This is always true when using a channel decoder.) The probability of error is $\sum_{w=1}^M \Pr(w, D_w^c)$. Define

$$\Omega = \{(w, b^T) : \frac{1}{T} \log \frac{P_{W,B^T}(w,b^T)}{P_W P_{B^T}(w,b^T)} \leq \frac{1}{T} \log M - \gamma\}$$

and

$$\Omega_w = \{b^T : \frac{1}{T} \log \frac{P_{W,B^T}(w,b^T)}{P_W P_{B^T}(w,b^T)} \leq \frac{1}{T} \log M - \gamma\}.$$

Now

$$\begin{aligned}
\Pr(w, D_w^c) &\geq \Pr(w, \Omega_w) - \Pr(w, \Omega_w \cap D_w) \\
&= \Pr(w, \Omega_w) - \sum_{b^T \in \Omega_w \cap D_w} P_{W,B^T}(w, b^T) \\
&= \Pr(w, \Omega_w) - \sum_{b^T \in \Omega_w \cap D_w} \frac{P_{WB^T}(w,b^T)}{P_W P_{B^T}(w,b^T)} P_W P_{B^T}(w, b^T)
\end{aligned}$$

$$\begin{aligned}
&\geq \Pr(w, \Omega_w) - \sum_{b^T \in \Omega_w \cap D_w} M e^{-T\gamma} \, P_W P_{B^T}(w, b^T) \\
&\geq \Pr(w, \Omega_w) - \sum_{b^T \in D_w} M e^{-T\gamma} \, P_W P_{B^T}(w, b^T) \\
&\geq \Pr(w, \Omega_w) - M e^{-T\gamma} P_W(w) \Pr(D_w) \\
&= \Pr(w, \Omega_w) - e^{-T\gamma} \Pr(D_w)
\end{aligned}$$

Thus

$$\begin{aligned}
\sum_{w=1}^{M} \Pr(w, D_w^c) &\geq \sum_{w=1}^{M} \left(\Pr(w, \Omega_w) - e^{-T\gamma} \Pr(D_w) \right) \\
&\geq \Pr(\Omega) - e^{-T\gamma} \sum_{w=1}^{M} \Pr(D_w) \\
&\geq \Pr(\Omega) - e^{-T\gamma}
\end{aligned}$$

Now by lemma 2.3 we may conclude

$$\epsilon \geq P_{A^T,B^T} \left(\frac{1}{T} \log \frac{P_{A^T,B^T}(a^T, b^T)}{\vec{P}_{A^T|B^T} P_{B^T}(a^T, b^T)} \leq \frac{1}{T} \log M - \gamma \right) - e^{-\gamma T} \qquad \forall \gamma > 0$$

□

Theorem 3.1 *The channel capacity $C^o \leq C$.*

Proof: Follows exactly as theorem 5 of [VH]. □

The converse theorem is really the easy part. This is due to the fact that any channel code induces a measure on the (A^T, B^T) space. But this measure need not be the measure that supremizes the directed mutual information. The direct theorem is harder to prove. We want to use a random coding argument. To that end we need to first generate a channel code from the channel input distribution. Then we need to show that at least one of these codes achieves capacity.

3.2 Direct Theorem

We will prove the direct theorem via a random coding argument. We construct a distribution, P_{F^T}, on $\mathcal{F}_T$ from the channel input distribution $\{P_{A_t|A^{t-1},B^{t-1}}\}$ via the following:

$$P(f_t | f^{t-1}) = \prod_{(b^{t-1}, a_t) \in \text{graph}(f_t)} P(a_t \mid f_1, f_2(b_1), \ldots, f_{t-1}(b^{t-1}), b^{t-1})$$

Let P_{F^T} be produced as above. By lemma 2.1 this P_{F^T} and the channel uniquely define a measure Q_{F^T,A^T,B^T}.

Lemma 3.2 *The channel input distribution induced from the Q measure equals the original channel input distribution. Specifically $Q_{A_t|A^{t-1},B^{t-1}} = P_{A_t|A^{t-1},B^{t-1}}$.*

Proof:

$$\begin{aligned}
Q_{A_t|A^{t-1},B^{t-1}}(a_t | a^{t-1}, b^{t-1}) &= Q_{F_t|A^{t-1},B^{t-1}} \left(\left\{ f_t : f_t(b^{t-1}) = a_t \right\} \mid a^{t-1}, b^{t-1} \right) \\
&= P_{A_t|A^{t-1},B^{t-1}}(a_t | a^{t-1}, b^{t-1})
\end{aligned}$$

□

3.2.1 Achievability

The following is a generalization of Feinstein's lemma [Fei].

Lemma 3.3 *Fix a time T and $0 < \epsilon < 1$. Then for all $\gamma > 0$ and channel input distributions $\{P_{A_t|A^{t-1},B^{t-1}}\}$, there exists an (T, M, ϵ) channel code for the channel $\{P_{B_t|B^{t-1},A^t}\}$ that satisfies*

$$\epsilon \leq P_{A^T,B^T}\left(\frac{1}{T}\log\frac{P_{A^T,B^T}(a^T,b^T)}{\vec{P}_{A^T|B^T}P_{B^T}(a^T,b^T)} \leq \frac{1}{T}\log M + \gamma\right) + e^{-\gamma T}.$$

Proof: We will bound the average probability of error over all codebooks chosen via the scheme mentioned above. Define

$$\Omega = \{(f^T, b^T) : \frac{1}{T}\log\frac{P_{F^T,B^T}(f^T,b^T)}{P_{F^T}P_{B^T}(f^T,b^T)} \leq \frac{1}{T}\log M + \gamma\}.$$

Recall the inequality $1-(1-x)^k \leq kx \quad \forall x \in (0,1)$. Then via the usual random coding arguments

$$\begin{aligned}
& E_{\text{codebooks}}(\epsilon) \\
\leq\ & P_{F^T,B^T}(\Omega) + \sum_{\Omega^c} \Pr(\text{error}|f^T,b^T)\ P_{F^T,B^T}(f^T,b^T) \\
=\ & \Pr(\Omega) + \sum_{\Omega^c} P_{F^T,B^T}(f^T,b^T)\left[1-\left(1-P_{F^T}\left\{\tilde{f}^T : \frac{P_{B^T|\tilde{F}^T}(b^T|\tilde{f}^T)}{P_{B^T|F^T}(b^T|f^T)} > 1\right\}\right)^{M-1}\right] \\
\leq\ & \Pr(\Omega) + \sum_{\Omega^c} P_{F^T,B^T}(f^T,b^T)(M-1)P_{F^T}\left\{\tilde{f}^T : \frac{P_{B^T|\tilde{F}^T}(b^T|\tilde{f}^T)}{P_{B^T|F^T}(b^T|f^T)} > 1\right\} \\
=\ & \Pr(\Omega) + (M-1)\sum_{\Omega^c} P_{F^T,B^T}(f^T,b^T)\left(\sum_{\tilde{f}^T}\frac{P_{B^T|\tilde{F}^T}(b^T|\tilde{f}^T)}{P_{B^T|F^T}(b^T|f^T)}P_{F^T}(\tilde{f}^T)\right) \\
=\ & \Pr(\Omega) + (M-1)\sum_{\Omega^c} P_{F^T,B^T}(f^T,b^T)\left(\frac{P_{B^T}(b^T)}{P_{B^T|F^T}(b^T|f^T)}\right) \\
\leq\ & \Pr(\Omega) + (M-1)\sum_{\Omega^c}\frac{1}{M}e^{-\gamma T}P_{F^T,B^T}(f^T,b^T) \\
\leq\ & \Pr(\Omega) + e^{-\gamma T}
\end{aligned}$$

Now by lemma 2.1 and lemma 3.2 we may conclude

$$E_{\text{codebooks}}(\epsilon) \leq P_{A^T,B^T}\left(\frac{1}{T}\log\frac{P_{A^T,B^T}(a^T,b^T)}{\vec{P}_{A^T|B^T}P_{B^T}(a^T,b^T)} \leq \frac{1}{T}\log M + \gamma\right) + e^{-\gamma T}.$$

Since the error averaged over codebooks is less than the right hand side we know there must exist at least one code with error

$$\epsilon \leq P_{A^T,B^T}\left(\frac{1}{T}\log\frac{P_{A^T,B^T}(a^T,b^T)}{\vec{P}_{A^T|B^T}P_{B^T}(a^T,b^T)} \leq \frac{1}{T}\log M + \gamma\right) + e^{-\gamma T}.$$

Thus the lemma is proved. □

Theorem 3.2 *All rates less than C are achievable.*

Proof: This follows from theorem 2 of [VH]. □

4 Channels with State

In general computing the capacity, C, is difficult. We examine a class of channels where one can use dynamic programming to compute the capacity. In this section we discuss channels with state. Let $\mathcal{Z}$ be a finite space representing the space where the state lives.

Definition 4.1 *A Markov channel is a channel such that the state evolves according to*

$$P_{Z_{t+1}|(A^t,B^t,Z^t)}(z_{t+1}|a^t,b^t,z^t) = P_{Z_{t+1}|A_t,Z_t}(z_{t+1}|a_t,z_t).$$

The channel output is modulated by the state in the following way

$$P_{B_t|(A^t,B^{t-1},Z^t)}(b_t|a^t,b^{t-1},z^t) = P_{B_t|A_t,Z_t}(b_t|a_t,z_t).$$

Note that the next state may depend on the current channel input. Thus our formulation contains ISI channels. We assume that the decoder observes both the state and the channel output. It then decides what to feedback to the encoder.

We define channel input distributions and the channel capacity just as in section 2.

Definition 4.2 *For all T let $\mathcal{S}_T = \{\{P_{a_t \mid z^t, a^{t-1}, b^{t-1}}\}_{t=1}^T\}$, $\mathcal{S}_T^{out} \subset \mathcal{S}_T$ be the set of all channel input distributions without state feedback, and $\mathcal{S}_T^{nfb} \subset \mathcal{S}_T^{out}$ be the set of all channel input distributions without any feedback.*

Definition 4.3 *For finite T let $C_T = \sup_{S\in\mathcal{S}_T} \frac{1}{T} I\left(A^T \to (Z^T, B^T)\right)$ and for the infinite horizon case let $C = \sup_{S\in\mathcal{S}_\infty} \underline{I}\,(A \to (Z,B))$. Define $C_T^{out}, C_T^{nfb}, C^{out}$, and C^{nfb} analogously.*

Just as in definition 2.8 we define the operational capacities under the different feedback scenarios: $C^{\mathrm{o}}, C^{\mathrm{o,\ out}}$, and $C^{\mathrm{o,\ nfb}}$.

4.1 Coding Theorem

We now prove the feedback Markov channel coding theorem. The Markov channel is a special case of the finite alphabet channel we examined in section 3. Thus we can apply theorems 3.1 and 3.2 directly.

Theorem 4.1 *$C^o = C$, $C^{o,\ out} = C^{out}$, and $C^{o,\ nfb} = C^{nfb}$.*

Proof: A Markov channel is just a special case of the channel we examined in section 3. Let $V_0 = Z_0$, $V_t = (B_t, Z_{t+1})$, $t = 1, ..., T-1$, and $V_T = B_T$. Then we can define a new "$A - V$"channel. Specifically this channel is $\{P(V_0),\ P(V_t \mid a^t, v^{t-1}), t = 1, ..., T\}$. There are two differences between this channel and the channels defined in section 2. First there is the addition of the initial state: $P(V_0)$. Second the spaces in which the V_t live in are changing with t. After straightforward corrections for these changes we can apply theorems 3.1 and 3.2 directly. □

4.2 The Optimization Problem

It can be shown that the directed mutual information can be written as the sum of "running costs," [Tat]:

$$I(A^T \to (B^T, Z^T)) = \sum_{t=1}^{T} I(A_t; B_t | B^{t-1}, Z^t) + \sum_{t=1}^{T-1} I(A_t; Z_{t+1} | B^t, Z^t)$$

Note that if there is no ISI the second sum is zero.

We prove one final simplification before presenting the details of the stochastic control problem. Define G_t to be the information available to the encoder at time t. Recall at one extreme: $G_t = (B^{t-1}, Z^t)$. One intermediate case is if there is no state feedback: $G_t = B^{t-1}$. And the other extreme occurs if there is no feedback at all: $G_t = \emptyset$. Let $\pi_t(g_t) = P(Z_t \,| g_t)$ be the conditional probability of the current state given the information available to the encoder at time t.

Lemma 4.1 *Π_t is a sufficient statistic.*

Proof: See [Tat]. □

4.2.1 The Dynamic Programming Equation

We now translate the elements of this problem into a traditional control problem.

- *State:* Let the state process be Z_t.
- *Control:* The control, U_t, takes values in $\mathcal{P}(\mathcal{A})$. It is allowed to be a function of the past controls and its observations G^t. At time t the controller draws an a_t from the distribution u_t and inputs it into the channel.
- *Observation:* The observation is G_t. The channel output follows $P(B_t \mid z_t, u_t) = \sum_{a_t} P(B_t \mid z_t, a_t) u_t(a_t)$.
- *Running and terminal cost* For $t = 1, ..., T-1$

$$c_t(z_t, u_t) = \sum_{a_t, b_t, z_{t+1}} P(b_t, z_{t+1} \mid z_t, a_t) u_t(a_t) \log \frac{P(b_t, z_{t+1} \mid z_t, a_t)}{\sum_{\tilde{a}_t} P(b_t, z_{t+1} \mid z_t, a_t) u_t(\tilde{a}_t)}$$

$$c_T(z_T, u_T) = \sum_{a_T, b_T} P(b_T \mid z_T, a_T) u_T(a_T) \log \frac{P(b_T | z_T, a_T)}{\sum_{\tilde{a}_T} P(b_T \mid z_T, a_T) u_T(\tilde{a}_T)}.$$

Lemma 4.2 *For the following cases $G_t = (B^{t-1}, Z^t)$ (full observation), $G_t = B^{t-1}$ (channel output feedback), and $G_t = \emptyset$ (no feedback at all) the process Π_t is a controlled Markov process.*

Proof: See [Tat] □.

We can now convert the above partially observed Markov decision problem into a fully observed Markov decision problem. The new state process is Π_t and the new running cost is

$$\bar{c}(\pi_t, u_t) = \sum_{z_t} c(z_t, u_t) \pi_t(z_t).$$

Theorem 4.2 *The optimal cost and control can be determined by solving the following dynamic programming equations*

$$J_T(\pi) = \sup_{u_T} \ \bar{c}(\pi, u_T) \quad \textit{and} \quad J_t(\pi) = \sup_{u_t} \ \bar{c}(\pi, u_t) + \int J_{t+1}(\tilde{\pi})P(d\tilde{\pi}|\pi, u_t)$$

where u_t is a function of π_t.

Proof: By lemmas 4.1 and 4.2 we have shown that Π_t is a sufficient statistic and a controlled Markov process. The theorem then follows from theorem 5.4.8 of [Str]. □

4.2.2 Infinite Horizon Average Cost Problem

So far we have dealt with the finite horizon problem. We can also treat the average cost infinite horizon problem: $\sup \liminf_{T\to\infty} \frac{1}{T} I(A^T \to (Z^T, B^T))$ where the supremum is over all admissible input distributions.

Assumption 4.1 *Assume that under all stationary control policies the process Z_t is an ergodic and aperiodic Markov chain.*

Note if there is no ISI then assumption 4.1 states that, independent of the policy, the Z_t process is an ergodic, aperiodic, Markov chain.

Theorem 4.3 *Under assumption 4.1 the optimal control is the solution to the following average cost Bellman equation*

$$J^* + h(\pi) = \max_u \ \bar{c}(\pi, u) + \int h(\tilde{\pi})P(d\tilde{\pi}|\pi, u)$$

Proof: See theorem 4.1 of [Bor]. □

References

[Bor] V. Borkar, "Average Cost Dynamic Programming Equations for Controlled Markov Chains with Partial Observations." To appear in the SIAM Journal on Control and Optimization.

[Fei] A. Feinstein, "A New Basic Theorem of Information Theory." IRE Trans. on Information Theory, Vol 4, pp. 2-22, 1954

[Mar] H. Marko, "The bidirectional Communication Theory - A Generalization of Information Theory." IEEE Trans. Commm., vol. COM-21, pp. 1345-1351, December 1973.

[Mas] J. Massey, "Causality, Feedback, and Directed Information." Proceedings of the 1990 International Symposium on Information Theory and its Applications (ISITA-90), pp. 303-305.

[Str] C. Striebel, *Optimal Control of Discrete Time Stochastic Systems.* Berlin, Springer-Verlag, 1975.

[Tat] S. Tatikonda, *Control Under Communication Constraints.* M.I.T. Ph.D. thesis, September 2000.

[VH] S. Verdu, T. Han, "A general formula for channel capacity." IEEE Trans. Info. Theory, vol. IT-40, pp. 1147 - 1157, July 1994.

Energy Efficiency, Energy Constraints, and Admission Control in Wireless Networks: A Case Study in Session Multicasting

Jeffrey E. Wieselthier and Gam D. Nguyen
Code 5521
Information Technology Division
Naval Research Laboratory
Washington, DC 20375
{wieselthier, nguyen} @ itd.nrl.navy.mil

Anthony Ephremides
Electrical and Computer Engineering Dept.
and Institute for Systems Research
University of Maryland
College Park, MD 20742
tony @ eng.umd.edu

ABSTRACT

In wireless networks, the constraints associated with *energy-limited* operation create a networking environment that differs from that of *energy-efficient* operation. In this paper, we discuss the similarities and differences between these two types of energy-aware scenarios, while focusing on the problem of multicasting of session-oriented traffic. We then address admission control from the perspective of pruning costly destinations from the multicast tree. Two approaches are presented, both of which exploit the properties of the wireless communication channel.

1. INTRODUCTION

We consider all-wireless (i.e., infrastructureless, peer-to-peer, or ad hoc) multihop networks, in which energy-related considerations are of primary importance. Alternative formulations can be developed depending on whether the objective is to provide energy-efficient operation, or to operate under limited-energy conditions. Energy-efficient operation refers to the economical use of energy, e.g., minimizing the total energy used per unit of traffic over an extended period of time. By contrast, energy-limited operation refers to the case in which the total energy available at each network node is limited as a consequence of the inability to recharge batteries during the network's lifetime. In such cases, the goal may be to provide acceptable performance for as long as possible. In this paper, we address both cases. In particular, we study the problem of source-initiated, multicast session traffic, although many of the trade-offs discussed in this paper apply to other applications as well.

It is important to recognize the differences between energy-limited and energy-efficient operation, and the conflicting but overlapping requirements that they impose on network operation. When a node is subject to a hard constraint on its total available energy, it will cease to function when its energy is depleted. By contrast, when the goal is energy efficiency, it is implicitly assumed that ample energy is available, but that it should be used economically; in such cases, energy use is factored in as a cost function. For example, use of the most energy-efficient multicast trees (hence energy-efficient operation) may result in premature depletion of energy at some nodes (and hence in inferior overall performance). The performance measures used to evaluate networking schemes should reflect the requirements associated with the specific application; e.g., some applications may tolerate the omission of some desired destinations from a session if doing so increases network lifetime.

In our earlier studies, we discussed the benefits of using "node-based" approaches for wireless networking applications, and we demonstrated how the characteristics of the wireless medium can be exploited. In [1] we developed the Multicast Incremental Power (MIP) algorithm for the construction of energy-efficient multicast trees, and in [2] we showed how it can be adapted for energy-limited applications as well. We showed that MIP performs considerably better than protocols that are based on "link-based" approaches, which are adaptations of protocols developed for wired networks, for a wide range of system parameters.

This work was supported by the Office of Naval Research.

In those studies, we assumed an "admit-all" admission-control policy, in which a tree is formed to reach all reachable destinations. A disadvantage of this policy is that the inclusion of "costly" destinations (i.e., those that require excessive energy to reach) in the trees can result in energy inefficiency and in the rapid depletion of energy at some nodes. The major contribution of this paper is the introduction of admission control schemes that exploit the properties of the wireless channel, and that provide both energy-efficiency and increased network lifetime under energy-constrained conditions. These admission-control schemes are suitable for use with MIP or, in fact, with any algorithm for multicast tree formation.

2. THE MODEL

We consider source-initiated, circuit-switched, multicast sessions. The maintenance of a session requires the dedication of a transceiver at each participating node (source node, relay nodes, and destination nodes) throughout the duration of the session. The network consists of N nodes, which are randomly distributed over a specified region. We assume in this paper that ample bandwidth is available, and that each node has a sufficient number of transceivers to accommodate all service requests. The constraints of finite bandwidth and transceiver resources can be accommodated as in [3], [4].

Any node is permitted to initiate multicast sessions. Multicast requests and session durations are generated randomly at the network nodes. Each multicast group consists of the source node plus at least one destination node. Additional nodes may be used as relays either to provide connectivity to all members of the multicast group or to reduce overall energy consumption. The set of nodes that support a multicast session (the source node, all destination nodes, and all relay nodes) is referred to as a *multicast tree*.

The connectivity of the network depends on the transmission power. We assume that each node can choose its power level p, such that $p_{min} \le p \le p_{max}$. The nodes in any particular multicast tree do not necessarily have to use the same power levels; moreover, a node may use different power levels for the various multicast trees in which it participates.

We assume that the received signal power is equal to $pr^{-\alpha}$, where p is the transmission power, r is the distance and α is a parameter that typically takes on a value between 2 and 4, depending on the characteristics of the communication medium. We use a simplified interference model in which we assume that the interference level is independent of network traffic and the same at all nodes. Based on this model, the transmitted power required to support a link between two nodes separated by distance r is proportional to r^{α}, since the received power must exceed some threshold. Without loss of generality, we set the threshold constant equal to 1, resulting in:

$$p_{ij} = \text{power needed for link between Node } i \text{ and Node } j = r_{ij}^{\alpha}, \qquad (1)$$

where r_{ij} is the distance between Node i and Node j.

We assume that energy is expended only for transmission, i.e., we neglect the energy needed for reception and signal processing.[1] We assume the use of omnidirectional antennas; thus all nodes within communication range of a transmitting node can receive its transmission. In such cases, we can exploit the "wireless multicast advantage," described in [1]. For example, consider a situation in which Node i transmits directly to its neighbors, Node j and Node k; the power required to reach Node j is P_{ij} and the power required to reach Node k is P_{ik}. A single transmission at power $P_{i,(j,k)} = \max\{P_{ij}, P_{ik}\}$ is sufficient to reach both Node j and Node k (rather than the sum of these powers, as in wired applications).

As a result of the wireless multicast advantage, the omnidirectional wireless communication medium can be viewed as a *node-based* environment. By contrast, in wired models, as long as there is a wire or cable link connecting two nodes, the reception is ensured over that link, and the cost of Node i's transmission to Node j and Node k would be the *sum* of the costs to the individual nodes. Thus, wired networks are *link-based.* The node-based nature of wireless networks necessitates the development of new networking techniques, because the models developed for wired networks do not adequately capture the characteristics of the wireless medium [1].

---

[1] It would be straightforward to incorporate receiver and signal processing energy into our model, but to do so would unnecessarily complicate our model at this stage.

The Broadcast Incremental Power (BIP) algorithm [1] is the first broadcasting algorithm to exploit the wireless multicast advantage. BIP is a centralized algorithm in which tree construction starts from the source node. One node is added to the tree at each step of construction until all nodes are included in the tree. BIP is similar to Prim's algorithm for the construction of minimum-cost spanning trees, but there is one crucial difference. Straightforward application of Prim's algorithm would involve finding the link of least cost (e.g., transmitter power) that would add a new node to the tree. By contrast, under BIP the node that can be added with least *incremental* cost is chosen. For example, if Node *i* is already transmitting with enough power to reach Node *j*, (i.e., with power r_{ij}^{α}) the incremental cost for Node *i* to reach Node *k* is the additional power needed ($r_{ik}^{\alpha} - r_{ij}^{\alpha}$). BIP is the basis of our MIP algorithm for multicasting, which is discussed in Section 4.

3. A MULTICASTING PROBLEM

We now address the problem of determining an appropriate multicast tree for each arriving multicast session request, so that a reward function (which incorporates throughput, while reflecting energy considerations) is maximized. The establishment of a multicast tree requires the specification of the transmitted power levels, the frequencies used by each node, and the commitment of the needed transceiver resources throughout the duration of the multicast session. Under our assumption of a sufficient number of transceivers and frequencies, a destination can be reached as long as there exists a path from the source to it (i.e., the transmitted power required to support the path does not exceed p_{max} at any node).

In our earlier studies, all multicast requests were accepted as long as one or more of the intended destinations can be reached, and paths are established to all reachable destinations, regardless of the cost required to do so. In this paper, we introduce a new method for admission control that eliminates costly destinations from the multicast tree.

3.1 Performance Measures

The suitability of a performance measure depends on whether the objective is energy-efficient or energy-limited operation. These objectives may be stated qualitatively as follows:

Energy-efficient operation:

- Maximize the total quantity of information per unit energy that is delivered to the destinations.

Energy-limited operation:

- Maintain operation at (or near) the best possible performance (e.g., reach most, if not all, destinations) for as long as possible.

In this paper, we limit our discussion to a single performance measure that is of interest for both of these cases. Since partial multicast sessions may take place (because some nodes may be unreachable), the performance metric should provide a reward that reflects the number of destinations that are actually reached. We define

n_i = # of intended destinations in session *i*;

m_i = # of destinations reached in session *i*;

P_i = sum of the transmitter powers used by all nodes in session *i*.

We define the *multicast efficiency* e_i of the *i*th multicast session to be the fraction of desired destinations of that session request that are actually reached: $e_i = m_i/n_i$. Then, the *cumulative multicast efficiency* over an observation interval of *X* multicast requests can be defined as:

$$e = \frac{1}{X}\sum_{i=1}^{X} e_i = \frac{1}{X}\sum_{i=1}^{X}\left(\frac{m_i}{n_i}\right). \tag{2}$$

In [15], to take into consideration the often-conflicting objectives of reaching as many destinations as possible and of doing so with the smallest amount of energy, we defined the *yardstick* Y, as follows:

$$Y = \frac{1}{X}\sum_{i=1}^{X}\left(\frac{m_i}{P_i}\right)\left(\frac{m_i}{n_i}\right). \tag{3}$$

In applications where energy efficiency is the goal, it is desirable to maximize the value of either e or Y, subject to a constraint on the average total power consumed by the network. By contrast, when there are hard constraints on the amount of energy at each node, one may be interested in either maintaining one of these performance measures at its highest possible value for as long as possible, or alternatively in achieving the highest possible value at the end of the network's lifetime (i.e., when the nodes have all died).

In [2] we also considered additional performance measures, including the total delivered traffic volume (which, for each session is proportional to the number of destinations reached and to the quantity of data in the session, and hence to session duration).

Tree formation consists of the choice of transmitting nodes and their transmitting powers. Our BIP and MIP algorithms (see [1]) uses node-based metrics in the construction of the tree, rather than the more-conventional link-based metrics. Link-based metrics assign a cost to the maintenance of each link, e.g., the power needed to maintain the link. The total cost of a multicast tree is then the sum of the costs of the links that form the tree. However, such metrics do not reflect the wireless multicast advantage property, discussed in Section 2. Instead, the total cost of a multicast tree should be evaluated as the sum of the costs of the transmitting nodes that form the tree.

4. ALGORITHMS FOR MULTICASTING

It is well known that the determination of a minimum-cost multicast tree in wired networks is a difficult problem, which can be modeled as the NP-complete Steiner tree problem, even though the broadcasting problem is easily formulated as the minimum-cost spanning tree (MST) problem, which has low complexity. The multicasting problem appears to be at least as hard in wireless networks as it is in wired networks, and we know of no scalable algorithms for the minimum-energy broadcast problem. Thus, heuristics are needed.

We have considered two basic approaches for multicasting, which are described in [1]:

- Pruning the broadcast tree produced by BIP:
  the Multicast Incremental Power (MIP) algorithm;
- Superposing the minimum-cost unicast paths to each individual destination:
  the Multicast Least-Unicast-cost (MLU) algorithm.

In [1] we evaluated these protocols, based on the criterion of energy-efficient multicast trees, and showed that MIP provides superior performance over a wide range of system parameters. In [3], [4], and [2] we evaluated these protocols based on operation over an extended period of time, and with finite transceiver and frequency resources. The admission-control schemes we propose in this paper can be used with any tree-construction protocol.

4.1 The Incorporation of Energy Limitations

If the cost metric used in the construction of the multicast trees does not reflect the constraint of finite energy at each node, the greedy nature of MIP and MLU (both of which construct trees without regard to the residual energy available at the nodes) can result in rapid energy depletion at some nodes. When nodes "die" in this manner, it may be no longer possible to create energy-efficient trees, and performance (in the sense of the chosen performance measure) can rapidly deteriorate.

We can discourage the inclusion of energy-depleted nodes in the multicast tree by increasing the cost associated with their use [2]. We define $E_i(t)$ to be the energy at Node i at time t. Thus, the initial value of energy at Node i is $E_i(0)$; we refer to the energy at time t as the *residual energy* at node i. We define the cost of a link between Nodes i and Node j to be

$$C_{ij} = p_{ij}\left(\frac{E_i(0)}{E_i(t)}\right)^{\beta}. \qquad (4)$$

where β is a parameter that reflects the importance we assign to the impact of residual energy. Clearly, when $\beta = 0$, the link cost is simply the power needed to maintain the link. Nodal costs are defined in a similar manner (see [2]):

$$C_i = p_i\left(\frac{E_i(0)}{E_i(t)}\right)^{\beta}. \qquad (5)$$

When β is too small, too much emphasis may be placed on the construction of energy-efficient trees, resulting in the depletion of energy at some of the nodes. By contrast, when β is too large, too much emphasis may be placed on balancing energy use throughout the network, while under-emphasizing the need for energy efficiency.

In [2] we showed that the incorporation of residual energy into local cost metrics, which results in spreading the burden of energy use among more of the nodes, has a considerable impact on network performance. Most significantly, we have shown that multicast efficiency can be maintained at high levels significantly longer when residual energy is taken into account; we have found that use of $\beta = 0.5$ typically provides significantly improved performance based on this criterion. On the other hand, the total data volume that can be delivered during the lifetime of the nodes is reduced slightly; in our examples, the total data volume is maximized when $\beta = 0$. Therefore, the benefit (or lack thereof) of taking residual energy into account depends strongly on the performance goals associated with the specific application.

5. ADMISSION CONTROL

We now introduce two admission-control schemes that exploit the wireless multicast advantage (see Section 2). Our goal is to prune costly destinations from the already formed multicast tree, thereby resulting in improved energy efficiency and in extended network lifetime when operating under the constraint of finite energy.

5.1 Incremental-Cost-Based Admission Control

Under the Incremental-Cost-Based admission-control (ICAC) scheme, a destination is pruned if its incremental cost exceeds a threshold value P_{thresh}, i.e., excluding this node from the tree will result in a saving of more than P_{thresh} units of power. In general, the quantity P_{thresh} does not have to be fixed a priori. For example, it may depend on the average power to reach destinations in the tree, on current levels of residual energy, or on priority levels of specific sessions or specific destinations. The ICAC scheme is best described by example. The steps of the admission-control scheme are outlined as follows, where we have set $P_{thresh} = 10$.

- **Construct a multicast tree**

We construct a multicast tree, e.g., by using MIP. An example of the pruning process for costly destinations for such a tree is shown in Fig. 1, which shows the Source and four destinations: D_1, D_2, D_3, and D_4. We have also labeled the "branching nodes:" *i*, *j*, and *k*.

- **Label link costs**

Label every link with the power needed to sustain the link (e.g., $p = r^{\alpha}$).[2]

- **Determine incremental link costs at branching nodes**

At each link exiting a "branching node" (*i*, *j*, and *k* in our example), define the "incremental link cost," which is the cost to maintain the link, *given that all other links exiting from the*

---

[2] We could also consider link costs that involve residual energy. An advantage of using only link power here is that the P_{thresh} threshold does not have to be changed as energy is depleted. Thus, one possible approach is to use residual energy considerations in the formation of the tree (as in Section 4.1), but to prune costly destinations based on power alone.

same branching node are in use:

Incremental link cost

= max{0, link cost – highest cost of any other link exiting the branching node}.

The incremental link costs are shown in parentheses. For example, consider the two links exiting Node *j*; their costs are 5 and 6, and their incremental costs are 0 and 1. The use of incremental costs in this manner enables us to exploit the wireless multicast advantage.

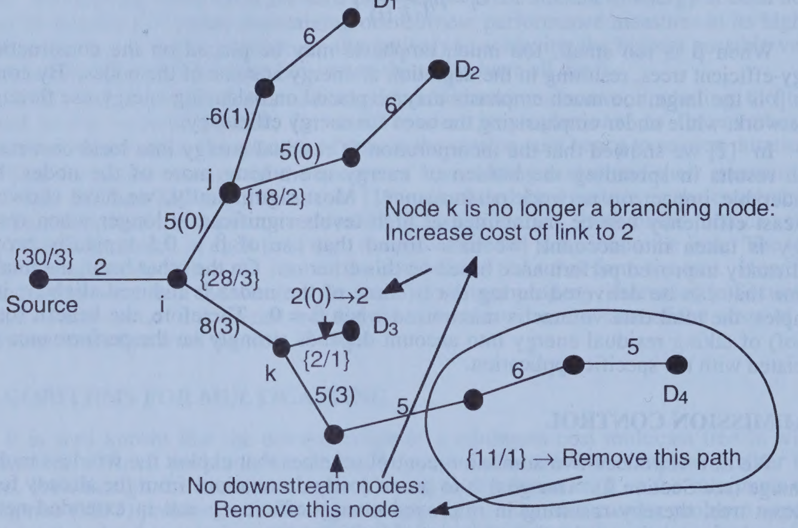

D1
6
D2
6
6(1)
5(0)
j
{18/2}
Node k is no longer a branching node:
Increase cost of link to 2
5(0)
{30/3}
2
{28/3}
Source
i
2(0)→2
8(3)
D3
{2/1}
k
5
6
5(3)
5
D4
{11/1} → Remove this path
No downstream nodes:
Remove this node

Fig. 1 — Example of Incremental-Cost-Based admission control, in which costly destinations are eliminated.

- **Delete paths to nodes with excessive incremental costs**

From each leaf destination node, go upstream one link at a time (stopping at a branching node):

- At each node:
  - Evaluate total cost on path to leaf node (= sum of link costs on all downstream links), using incremental link cost for links exiting branching nodes.
  - Evaluate {*A*/*B*}

    where *A* = total cost of downstream path to final leaf node(s)

    and *B* = number of destinations reached along this path

    and label branching points with this pair.

  If the ratio *A*/*B* is greater than P_{thresh}, delete the corresponding sub-path.

  Also delete any upstream nodes that no longer have any downstream destinations.

  (Example: see how path leading to D_4 is deleted.)

- **Update incremental link costs**

Whenever nodes are deleted, it may be necessary to change the incremental link costs at links exiting the same branching node. In our example, when the path to D_4 is removed, Node *k* is no longer a branching node, and the cost of the link from *k* to D_3 becomes 2, its full (i.e., not incremental) cost.

- **Prune costly sub-subtrees**
  - At each branching node:

    Evaluate $\{A/B\}$, where this quantity includes all paths exiting the branching node.

    The cost of the branching node's transmission is the max of all link costs.
  - If $\{A/B\}$ does not exceed P_{thresh}, keep the entire subtree rooted at the branching node.
  - If $\{A/B\}$ exceeds P_{thresh}, then evaluate the quantity $\{A/B\}$ for each sub-subtree that exits the branching node (i.e., the subtree that starts with a single link exiting the branching node, using the full [not incremental] cost of the link).

    Keep those sub-subtrees for which $\{A/B\} \leq P_{thresh}$.

    Discard those sub-subtrees for which $\{A/B\} > P_{thresh}$.

Note: If all combinations of sub-subtrees emanating from each branching point were examined, it might be possible to find additional costly nodes that should be pruned. But, doing so would increase complexity significantly.

- **Progress Upstream**

Once $\{A/B\}$ is evaluated at a branching node, this value is used as we progress upstream (i.e., we don't have to evaluate the individual downstream path costs anymore).

- **Observations**

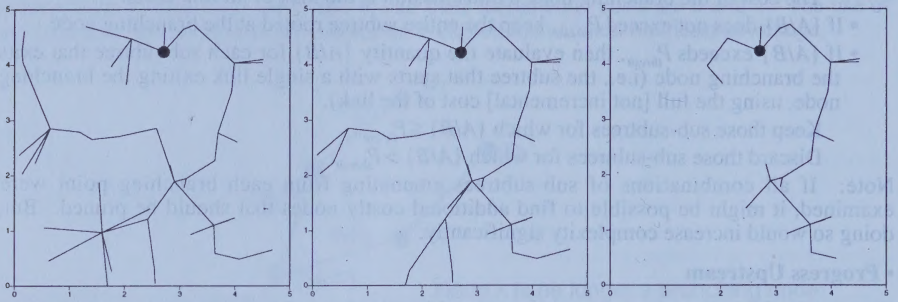
Now let us consider an implementation in which total link cost (rather than incremental link cost) is used to make the decision on pruning. For example, let us consider the subtree rooted at Node *j*. The total costs to reach D_1 and D_2 are 12 (6 + 6) and 11 (5 + 6), respectively. Thus, both of these destinations would be eliminated from the tree. Now, consider the cost of the path from the Source to D_3, which is 12 (2 + 8 + 2); thus D_3 would also be eliminated. Therefore, all destinations would be eliminated, even though exploitation of the wireless multicast advantage permits D_1, D_2, and D_3 to be included in the tree for a total cost of 30.

An advantage of the Incremental-Cost-Based admission control scheme is that it can be executed distributedly. Execution can be started from any leaf node, a node performs its calculations when it has heard from all downstream neighbors, and the decision on whether or not to prune a subtree is independent of that for any other disjoint subtree. However, a disadvantage is that it may be difficult to choose an appropriate threshold for acceptable values of incremental cost. One possible approach is to choose the threshold value for each multicast request, with the goal of excluding destinations whose incremental cost is greater than some multiple of the average cost per destination. An appropriate value for this multiple may depend on the residual energy at the nodes, and thus may be time-varying. Alternatively, the threshold may be based on the recent history of network operation, rather than only on the tree associated with the current multicast session request.

- **Examples**

Figure 2(a) shows the broadcast tree produced by BIP for a 50-node network, in which the source node is shown as a large circle. Figures 2(b) and 2(c) show the trees that result from the application of the ICAC algorithm with an incremental power threshold values of $P_{thresh} = 0.4$ and 0.25, respectively. For each of these trees, we show *d* (the number of destinations reached), P_{tree} (the total power needed to maintain the tree; we assume $\alpha = 2$, which results in $p_{ij} = r_{ij}^2$ as in Eq. (1)), and the ratio d/P_{tree}. As P_{thresh} is decreased, the number of destinations reached decreases, and the value of d/P_{tree} increases slightly.

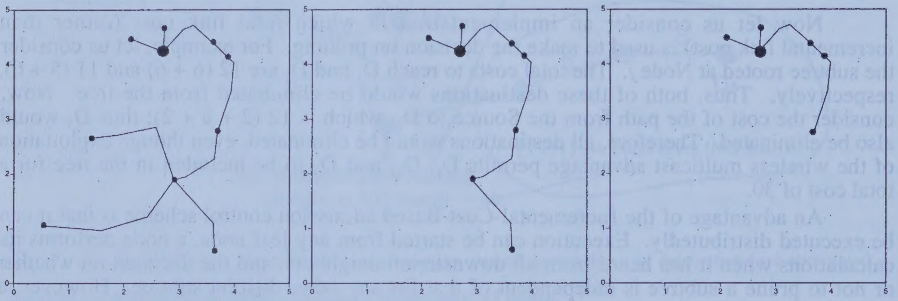
We now consider the case of multicasting for the same 50-node network, with the same source node and a multicast group size of 10. Figure 3(a) shows the resulting tree for the admit-all policy, and Figs. 3(b) and 3(c) show the effect of using ICAC thresholds of $P_{thresh} = 1.2$ and 1.0, respectively. The destination nodes are highlighted. In this example, the d/P_{tree} ratio more than doubles when the threshold P_{thresh} is decreased from ∞ to 1.0. Thus, there is a significant increase in the number of destinations reached per unit power, unlike the case of broadcasting discussed above. It is not surprising that the increase in the d/P_{tree} ratio is much greater than in the broadcasting case. In the multicasting case, since not all tree members are destinations (i.e., some are relays), there are more opportunities to remove nodes that do not contribute to the value of d (but, of course, do contribute to the cost of reaching destination nodes).



$d = 49$; $P_{tree} = 10.18$; $d/P_{tree} = 4.81$ $d = 42$; $P_{tree} = 8.32$; $d/P_{tree} = 5.05$ $d = 27$; $P_{tree} = 4.58$; $d/P_{tree} = 5.90$

(a) $P_{thresh} = \infty$ (b) $P_{thresh} = 0.4$ (c) $P_{thresh} = 0.25$

Fig. 2 — Application of ICAC admission control to 50-node broadcast tree produced by BIP.



$d = 9$; $P_{tree} = 6.39$; $d/P_{tree} = 1.41$ $d = 7$; $P_{tree} = 3.81$; $d/P_{tree} = 1.84$ $d = 5$; $P_{tree} = 1.73$; $d/P_{tree} = 2.89$

(a) $P_{thresh} = \infty$ (b) $P_{thresh} = 1.2$ (c) $P_{thresh} = 1.0$

Fig. 3 — Application of ICAC admission control to 50-node 10-destination tree produced by MIP.

We now consider the operation of a multicast network over a period of time. We have simulated the performance of MIP for a network of 100 nodes that are randomly located in a region with dimensions 5 × 5 (arbitrary units of distance), similar to that shown above. Also, as above, we have obtained results for a propagation constant value of $\alpha = 2$. In our simulations, multicast requests arrive with interarrival times that are exponentially distributed with rate λ/N at each node; we have used $\lambda = 1$ in our simulations. Session durations are exponentially distributed with mean 1. Multicast groups are chosen randomly for each session request; the number of destinations is uniformly distributed between 1 and N–1. In this paper, we assume that there are sufficient transceivers at each node and frequencies available to support trees to all destinations. Thus, there is no congestion, and destinations are excluded only on the basis of the admission-control algorithm. Consequently, we are able to assess the impact of admission control independently of congestion-related effects. Each simulation run consists of X = 1,000 multicast session requests. Table 1 shows the effect of varying P_{thresh} on several performance indexes, the yardstick Y (see Section 3.1), the average tree power $\bar{P}_{tree}$, the multicast efficiency e, and the ratio of the last two quantities.

With the admit-all policy (i.e., $P_{thresh} = \infty$), we of course have $e = 1.0$, since all destinations are always included in the tree, as noted above. As P_{thresh} is lowered, the number of destinations reached decreases, resulting in corresponding decreases in e and $\bar{P}_{tree}$. The highest value of Y is observed for $P_{thresh} = 0.5$. The $e/\bar{P}_{tree}$ ratio doesn't vary much for P_{thresh} >

0.5; however, when P_{thresh} is decreased to values lower than 0.5 this ratio increases noticeably, reflecting the fact that a significant number of costly destinations have been pruned.

Table 1 — Performance measures as a function of admission-control threshold P_{thresh}.

P_{thresh}	Y	$\overline{P}_{tree}$	e	$e/\overline{P}_{tree}$
0.1	2.277	1.07	0.169	0.158
0.15	4.622	3.46	0.463	0.134
0.20	5.417	4.79	0.621	0.130
0.25	5.732	5.57	0.713	0.128
0.5	5.899	7.18	0.886	0.123
0.75	5.783	7.82	0.939	0.120
1.0	5.774	8.04	0.967	0.120
∞	5.772	8.22	1.000	0.122

5.2 Yardstick-Based Admission Control

We now introduce an approach that does not have the shortcoming associated with an arbitrary admission-control threshold. In Section 3.1 we introduced the Yardstick measure Y, which incorporates both of our primary objectives, namely reaching many of the desired destinations, while conserving energy. Yardstick-based admission control tries to find the multicast tree that provides the largest value of Y. The problem of finding the multicast tree with the largest value of Y (for any particular source and set of desired destinations) is intractable; we know of no alternatives to exhaustive search. To reduce the complexity, we limit our search to the set of trees that can be formed by pruning a multicast tree that reaches all of the desired destinations (e.g., the tree produced by MIP). Nevertheless, despite this simplification, the complexity remains high and heuristics are needed to reduce the search space further.

The main idea is to compute the yardsticks associated with different trees, and then choose the tree that has the highest yardstick value. To keep the computational complexity low, we limit the search to a rather small number of trees; i.e., we would like the size of the search space, to be $O(N)$, where N is the number of nodes in the network. In the following, we describe two simple schemes to specify the search space. These schemes can serve as the basis for improved schemes; no claims of nearly optimal performance are made.

YS1: Subtrees produced in broadcast tree construction process

The search space consists of the subtrees produced by the steps in some tree-construction algorithm. For example, if the algorithm chosen is BIP, then the search space consists of N subtrees. The first subtree consists of only the source node (with zero yardstick); the second subtree consists of the source node and the nearest node to the source; and so on. The N^{th} subtree is the broadcast tree formed by BIP. For each of these N subtrees, we first prune nodes that have no downstream destinations, and then compute the value of the yardstick of the resulting tree. Finally, we choose the subtree that produces the largest yardstick value.

YS2: Subtrees produced by pruning the multicast tree

In this scheme, we are given an initial tree (which is produced by any tree-construction algorithm or by YS1, discussed in the preceding paragraph). Then the search space consists of the subtrees, each of which is formed by deleting a node (as well as all of its descendants). Thus, the i^{th} subtree is the tree with Node i and its descendants deleted; all other nodes still remain on the subtree (when Node i is the source, the subtree is empty and has zero yardstick). If the initial tree has N nodes, we can form N such subtrees. As above, we choose the subtree of maximum yardstick.

Table 2 summarizes the performance of our simulation runs of X = 1,000 multicast sessions, using the same scenario discussed above for the ICAC scheme. Results are presented for YS1, YS2, and for the concatenation of these two schemes, which we denote by YS1&2. The combined scheme results in a higher value of average yardstick, along with lower value of average tree power, but results in the lowest value of multicast efficiency e.

Table 2 — Performance measures for three alternative yardstick-based admission-control schemes.

Scheme	Y	$\overline{P}_{tree}$	e	$e/\overline{P}_{tree}$
YS1	6.095	7.57	0.947	0.125
YS2	6.031	7.58	0.961	0.127
YS1&2	6.193	7.21	0.932	0.129

6. CONCLUSIONS

In this paper, we have addressed the similarities and differences between energy-limited and energy-efficient operation. When the energy available for transmission at each node is limited, even the definition of performance measures becomes a difficult question to address, and the relative performance of multicasting schemes depends strongly on the criteria used to evaluate them. We showed in [2] that the incorporation of residual energy into local cost metrics, which results in spreading the burden of energy use among more of the nodes, has a considerable impact on network performance. The benefit (or lack thereof) of taking residual energy into account depends strongly on the performance goals associated with the specific application.

We have introduced novel admission-control schemes for wireless multicasting, in which "costly" destinations (i.e., those requiring excessive energy to reach) are pruned from the multicast tree. These schemes, which are the first that we know of that exploit the wireless multicast advantage property, provide enhanced energy efficiency and increase the useful lifetime of energy-limited networks. Our approach permits the trade-off between system performance and operational costs, including issues such as multicast efficiency *e*, useful network operational lifetime (when energy is constrained), cost of energy (e.g., price and weight of batteries), and priority mechanisms for different classes of traffic (or destinations).

REFERENCES

[1] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, "On the Construction of Energy-Efficient Broadcast and Multicast Trees in Wireless Networks," *Proceedings of IEEE INFOCOM 2000*, pp. 585-594, March 2000.

[2] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, "The Conundrum of Energy Efficiency in Energy-Limited Wireless Networks for Session-Based Multicasting," *submitted to IEEE INFOCOM 2001*, to take place April 2001.

[3] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, "Algorithms for Bandwidth-Limited Energy-Efficient Wireless Broadcasting and Multicasting," *Proceedings of MILCOM 2000*, October 2000.

[4] J. E. Wieselthier, G. D. Nguyen, and A. Ephremides, "Energy-Efficient Wireless Multicast of Session Traffic," to appear in Proc. *Hawaii International Conference on System Sciences (HICSS-34)*, to take place January 2001.

Admission Control Schemes to Provide Class-level QoS in Multiservice Networks*

Suresh Kalyanasundaram[†] Edwin K. P. Chong
Ness B. Shroff
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
E-mail:{kalyanas, echong, shroff}@ecn.purdue.edu

Abstract

In this work, we consider the scenario where different classes of traffic with differing bandwidth requirements are to be supported over a link with finite capacity. The network will be required to meet class-level QoS constraints that are specified as relations between the call blocking probabilities of these traffic classes. The network will also be required to maintain maximum throughput performance. We develop two different admission control schemes that consider these two important criteria. First, we choose an objective that maximizes the throughput subject to the constraint that the maximum deviation from the desired class-level QoS is smaller than a certain pre-specified quantity. In the second approach, we choose an objective that minimizes the deviation from the desired class-level QoS subject to the throughput being greater than a pre-specified minimum value. We show how these problems can be converted to linear programming problems. We illustrate our schemes using numerical examples.

1 Introduction

The trend of future networks, whether wired or wireless, is towards multiservice networks, where a single network infrastructure will support a variety of applications such as data, voice, and video [1]. Thus future networks will handle traffic belonging to different classes that have different quality-of-service (QoS) and bandwidth requirements. In this work, we are concerned with the call blocking probability requirement of different classes of users. The scenario we consider is one where calls belonging to several traffic classes compete for bandwidth on a link with finite capacity. All calls belonging to one traffic class are assumed to require the same amount of bandwidth. We also assume that users that are not admitted immediately are lost, i.e., we do not assume call queueing.

The assumption that each user requires a certain amount of bandwidth is clearly applicable to circuit-switched networks, e.g., circuit-switched wireless and optical networks. However, it does not restrict us to the case of circuit-switched networks alone.

---

*This research was supported in part by the National Science Foundation through grants NCR-9624525, CDA-9617388, ECS-9501652, and ANI-9805441, and by DARPA/ITO through grant F19628-98-C-0051

[†] Currently with Motorola Labs, Schaumburg, IL. E-mail: csk090@email.mot.com.

For packet-switched networks, although there is no physical reservation of bandwidth for each user, a logical notion of bandwidth requirement is still essential in providing QoS [2]. In packet-switched networks with statistical QoS guarantees, the amount of bandwidth needed by a user would be given by the effective or equivalent bandwidth [3, 4]. The above view of abstracting the bandwidth or rate requirement of a user hides the packet-level details, and simplifies the analysis. A justification for such a decoupling can be found in [5].

When a new call arrives, a *call admission control* decision occurs in which a controller decides whether or not the call can be admitted based on the current "state" of the system and the overall objectives of the network. The call admission control problem has been well-studied under different contexts. A common call admission policy is the *complete sharing* policy [6, 7, 8, 9]. Under this scheme, all the available link capacity is open to all classes of users. The *complete partitioning* [7, 8, 9] scheme, on the other hand, divides the entire capacity into as many partitions as there are classes of users, and a new user that finds insufficient capacity in the partition for its class is dropped. Other admission control schemes have been developed that meet specific objectives. For example, in [7] and [10], the authors formulate the problem of finding an admission control strategy that maximizes the throughput as a continuous-time Markov decision process. The authors of [11] are interested in making all classes of users experience the same call blocking probability.

Several future developments are likely to require call admission schemes to meet *class-level QoS* requirements. By class-level QoS, we mean the desired relations between the call blocking probabilities of different call classes. It is foreseen that appropriate pricing mechanisms will be introduced to prevent congestion in future networks [12]. The network will offer various levels of service with appropriate prices associated with them. All users who pay identical amounts for the same application belong to the same class. Clearly, users who pay more should be accorded priority and should belong to a higher priority class. The call blocking probability of high priority classes should be smaller than that of low priority classes. Such a requirement on the call blocking probability of call classes is an example of a class-level QoS requirement. Another example of a class-level QoS requirement is "fair blocking," where it is desired that all call classes experience the same call blocking probability. We have a very general formulation of class-level QoS requirements that includes the above examples as special cases. In our work, we consider relations of a special kind that achieve "weighted-fair blocking." Besides class-level QoS, throughput efficiency is another important factor to consider while designing call admission schemes. In this work, we design call admission schemes that consider these two criteria.

2 Maximizing Throughput Subject to a Class-level QoS Constraint

In this section, we describe a call admission scheme that maximizes throughput subject to a constraint on the maximum deviation from the desired class-level QoS relations. Let the number of call classes be M. The bandwidth requirement of calls in each class is fixed, but the requirement may vary from class to class. Let the bandwidth requirement of a class-i user be e_i bandwidth units (BWUs), and let the capacity of the link be C BWUs. Let calls of class i, $i = 1, 2, \ldots, M$, arrive according to an independent Poisson process with rate λ_i calls/sec, and let the call holding duration of class-i calls be independent and exponentially distributed with mean $1/\mu_i$ seconds. Depending on the characteristics of the call classes, their call blocking probabilities will be expected to satisfy several conditions. For example, in fair blocking the objective is to ensure that the difference in call blocking probabilities of any two different call classes is small. In other words, the desired relation among the call blocking probabilities of call classes is $B_1 = B_2 = \cdots = B_M$, where B_i denotes the call blocking probability of class-i users. In other situations, it would be necessary to give preferential treatment to high-priority call classes. To handle all such situations, we consider the general case of "weighted-fair blocking." (Although we consider the case of weighted-fair blocking in this work, our methods apply in the general setting of any linear relation involving the call blocking probabilities. We restrict ourselves to the case of weighted-fair blocking for ease of exposition.) In weighted-fair blocking, the objective is to reduce the difference between appropriately scaled call blocking probabilities. Thus, the desired relation between call blocking probabilities of classes is given by $w_{ls}B_l = w_{sl}B_s$, $1 \le l \le M$, $1 \le s \le M$, $l \ne s$ for appropriately chosen positive weights w_{ls}. A requirement that class-1 users experience half the call blocking probability of class-2 users is an example of weighted fair blocking, where we give preferential treatment to class-1 users. We note that there need not exist desired relations between the call blocking probabilities of every pair of classes, which is a special case of what we are considering. Another special case of our weighted-fair blocking is the case of bounds on the maximum call blocking probability of each call class. Also, the weights w_{ls} need to be chosen so that they do not result in inconsistent requirements on the B_l. If the chosen w_{ls} results in inconsistent requirements on the B_l, our method will reveal that it is so.

We denote the state of the system by a vector of M components $(n_1, n_2, \ldots, n_M)$ such that n_i represents the number of active class-i users in the system. We define the set S of all feasible states as

$$S = \left\{ \boldsymbol{n} = (n_1, n_2, \ldots, n_M) : \sum_{i=1}^{M} n_i e_i \le C \text{ and } n_i \in \mathbb{Z}^+ \text{ for } i = 1, 2, \ldots, M \right\},$$

where $\mathbb{Z}^+$ is the set of all non-negative integers. The set S includes all possible states such that the combined bandwidth requirement of the active users does not exceed the link bandwidth.

Given the current system state, the admission controller makes an admission control decision on the classes of calls that will be admitted in that state. In each state, the admission controller can choose one out of several possible actions. Formally, an action is denoted by a vector $\boldsymbol{u}$ such that $\boldsymbol{u} = (u_1, u_2, \ldots, u_M)$ with $u_i = 0$ or 1. If $u_i = 0$ then the admission controller blocks any call belonging to class i, and if $u_i = 1$ it accepts any call belonging to class i. We define the set $K_{\boldsymbol{n}}$ of all allowed control actions in state $\boldsymbol{n}$ as $K_{\boldsymbol{n}} = \{(u_1, u_2, \ldots, u_M) : u_i = 0 \text{ or } 1 \text{ for } i = 1, 2, \ldots, M \text{ and } \boldsymbol{n} + u_i \boldsymbol{\delta}_i \in S \text{ for } i = 1, 2, \ldots, M\}$, where $\boldsymbol{\delta}_i$ is a vector of M components with zeros in all except the i-th position where it has a one. Note that an action $\boldsymbol{u}$ is allowed in state $\boldsymbol{n}$ only if there is enough idle capacity in the system to admit a call belonging to any of the traffic classes that the action $\boldsymbol{u}$ decides to admit. The set $K_{\boldsymbol{n}}$ contains precisely these actions. For each state $\boldsymbol{n} \in S$, the set $K_{\boldsymbol{n}}$ has at most 2^M elements. A *control policy* is a sequence of such admission control actions made in each state as the system evolves. We let $\mathcal{F}'$ denote the set of all such control policies. Our objective is to obtain an admission control policy that optimizes the following problem:

$$\underset{f \in \mathcal{F}'}{\text{maximize}}\ T(f)$$

$$\text{subject to } \max\{|w_{ls}B_l(f) - w_{sl}B_s(f)| : 1 \leq l < M, l+1 \leq s \leq M\} \leq \epsilon, \tag{1}$$

where $0 < \epsilon < 1$ is a pre-specified constant, and $T(f)$ is the throughput (time-average of bandwidth usage) obtained using the policy f. The value of ϵ can be set to be the maximum deviation from the desired class-level QoS that can be tolerated. Clearly, the throughput and the call blocking probability of each call class depends on the chosen control policy $f \in \mathcal{F}'$. We show this explicitly in the problem formulation above. Equation (1) prevents a deviation larger than the allowed tolerance of ϵ from the desired relation between call blocking probabilities. As mentioned earlier, we consider the case of weighted-fair blocking, where the desired relation between call blocking probabilities is $w_{ls}B_l = w_{sl}B_s$ for $1 \leq l \leq M$, $1 \leq s \leq M$, $l \neq s$, and we obtain the case of fair blocking by setting $w_{ls} = 1$ for all l and s. We rewrite the constraint in Equation (1) as the following $M(M-1)$ constraints:

$$w_{ls}B_l - w_{sl}B_s \leq \epsilon \text{ for } 1 \leq l \leq M, 1 \leq s \leq M, \text{ and } l \neq s. \tag{2}$$

We now identify a per-unit-time reward such that the long-run average reward per unit time represents the throughput obtained from the chosen control policy $f \in \mathcal{F}'$. When the system stays in state $\boldsymbol{n} = (n_1, n_2, \ldots, n_M)$ and when action $\boldsymbol{u}$ is chosen, the reward is accrued at the rate of

$$r_{\boldsymbol{n}}^{\boldsymbol{u}} = \sum_{i=1}^{M} n_i e_i \text{ for all } \boldsymbol{u} \in K_{\boldsymbol{n}}, \tag{3}$$

where $K_{\boldsymbol{n}}$ is the set of all allowed control actions in state $\boldsymbol{n}$. The reward $r_{\boldsymbol{n}}^{\boldsymbol{u}}$ in state $\boldsymbol{n}$ is the amount of bandwidth in use in that state.

Besides the reward $r_{\boldsymbol{n}}^{\boldsymbol{u}}$, we now associate costs with each state $\boldsymbol{n}$ and for each action $\boldsymbol{u} \in K_{\boldsymbol{n}}$ such that the long-run average cost is the term on the left hand side of Equation (2). To do this, note that we can obtain the term B_l as the long-run average, if we associate a per-unit-time cost of one with states that choose an action that rejects class-l calls and a per-unit-time cost of zero with states that choose an action that admits class-l calls. The long-run average of this cost definition is the proportion of time the system spends in a state that rejects class-l calls. From this observation, it follows that the following per-unit-time cost definition will yield the term $w_{ls}B_l - w_{sl}B_s$ as the long-run average:

$$c_{\boldsymbol{n}}^{\boldsymbol{u}}(l,s) = \begin{cases} 0 & \text{if } \boldsymbol{u}_l = \boldsymbol{u}_s = 1 \\ w_{ls} & \text{if } \boldsymbol{u}_l = 0 \text{ and } \boldsymbol{u}_s = 1 \\ -w_{sl} & \text{if } \boldsymbol{u}_l = 1 \text{ and } \boldsymbol{u}_s = 0 \\ w_{ls} - w_{sl} & \text{if } \boldsymbol{u}_l = \boldsymbol{u}_s = 0. \end{cases} \tag{4}$$

If the system is in state $\boldsymbol{n}$ and action $\boldsymbol{u} \in K_{\boldsymbol{n}}$ is chosen, then the next state of the system is determined according to the rates $q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}}$, and these rates depend only on the current state of the system $\boldsymbol{n}$ and the chosen control action $\boldsymbol{u}$. The transition rates $q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}}$ are obtained as follows:

$$q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}} = \begin{cases} \lambda_i & \text{if } \boldsymbol{p} = \boldsymbol{n} + \boldsymbol{\delta}_i \text{ and } \boldsymbol{u}_i = 1 \\ n_i \mu_i & \text{if } \boldsymbol{p} = \boldsymbol{n} - \boldsymbol{\delta}_i \text{ and } n_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

From the above discussion it follows that we have a constrained continuous-time Markov decision problem to solve (see [13, 14, 15]). It can be shown that we can search for the optimal policy over the smaller set of stationary policies (policies that are independent of the history of the system and the time of decision making) instead of the original set $\mathcal{F}'$. To characterize the set $\mathcal{F}$ of all stationary policies we denote the probability that we choose control action $\boldsymbol{u} \in K_{\boldsymbol{n}}$ when the system is in state $\boldsymbol{n} \in S$ as $d_{\boldsymbol{n}}^{\boldsymbol{u}}$. The $d_{\boldsymbol{n}}^{\boldsymbol{u}}$ for $\boldsymbol{u} \in K_{\boldsymbol{n}}$ and $\boldsymbol{n} \in S$ are the decision variables of the problem. It is clear that

$$\sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} d_{\boldsymbol{n}}^{\boldsymbol{u}} = 1 \text{ for } \boldsymbol{n} \in S \tag{5}$$

and that

$$d_{\boldsymbol{n}}^{\boldsymbol{u}} \geq 0 \text{ for } \boldsymbol{u} \in K_{\boldsymbol{n}} \text{ and } \boldsymbol{n} \in S. \tag{6}$$

A stationary control policy is the following set of probability distributions, one for each feasible state:

$$\{d_{\boldsymbol{n}}^{\boldsymbol{u}} : \boldsymbol{u} \in K_{\boldsymbol{n}} \text{ and } \boldsymbol{n} \in S\}. \tag{7}$$

Implicit in the above notation is that the control policy does not depend on the time of decision making. In other words, successive visits to a particular state will employ the same probability distribution to choose the control action. The set $\mathcal{F}$ consists of distributions of the form given in Equation (7) satisfying Equations (5) and (6). Based on the above definition of a stationary control policy, we have the following definitions for pure and randomized policies. A stationary control policy is called a *pure policy* if for each $\boldsymbol{n} \in S$ there is an action $\boldsymbol{u} \in K_{\boldsymbol{n}}$ such that $d_{\boldsymbol{n}}^{\boldsymbol{u}} = 1$. A stationary policy that is not pure is a *randomized policy.*

Under any given stationary policy f, the states the system visits form a continuous-time Markov chain with transition rates given as follows:

$$q_{(\boldsymbol{n},\boldsymbol{p})}(f) = \sum_{\{\boldsymbol{u}:\boldsymbol{u} \in K_{\boldsymbol{n}}\}} q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}} d_{\boldsymbol{n}}^{\boldsymbol{u}}. \tag{8}$$

We let $\pi_{\boldsymbol{n}}(f)$ be the limiting steady-state probability of the continuous-time Markov chain for the stationary policy f. We know that these probabilities have to satisfy the following conditions:

$$\pi_{\boldsymbol{n}}(f) \sum_{\boldsymbol{p} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}} d_{\boldsymbol{n}}^{\boldsymbol{u}} - \sum_{\boldsymbol{i} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{i}}} \pi_{\boldsymbol{i}}(f) q_{(\boldsymbol{i},\boldsymbol{n})}^{\boldsymbol{u}} d_{\boldsymbol{i}}^{\boldsymbol{u}} = 0 \text{ for } \boldsymbol{n} \in S \tag{9}$$

$$\sum_{\boldsymbol{n} \in S} \pi_{\boldsymbol{n}}(f) = 1$$

$$\pi_{\boldsymbol{n}}(f) \geq 0 \text{ for } \boldsymbol{n} \in S.$$

Equation (9) is the balance equation for the continuous-time Markov chain and the interpretation in Equation (8) is useful in seeing that Equation (9) is valid. With the above definition of $\pi_{\boldsymbol{n}}(f)$, the long-run average reward reduces to $\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} \pi_{\boldsymbol{n}}(f) r_{\boldsymbol{n}}^{\boldsymbol{u}} d_{\boldsymbol{n}}^{\boldsymbol{u}}$, and is independent of the initial state of the system [16]: The product of $\pi_{\boldsymbol{n}}(f)$ and $d_{\boldsymbol{n}}^{\boldsymbol{u}}$ is the fraction of time the system is in state $\boldsymbol{n}$ and action $\boldsymbol{u}$ is chosen. The long-run average reward is merely the weighted average reward with weights being the product of $\pi_{\boldsymbol{n}}(f)$ and $d_{\boldsymbol{n}}^{\boldsymbol{u}}$. Similarly, the constraints on $w_{ls}B_l - w_{sl}B_s$ can be written as follows:

$$\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} \pi_{\boldsymbol{n}}(f) c_{\boldsymbol{n}}^{\boldsymbol{u}}(l,s) d_{\boldsymbol{n}}^{\boldsymbol{u}} \leq \epsilon \text{ for } 1 \leq l \leq M, 1 \leq s \leq M,\ l \neq s.$$

Setting $x_{\boldsymbol{n}}^{\boldsymbol{u}} = \pi_{\boldsymbol{n}}(f) d_{\boldsymbol{n}}^{\boldsymbol{u}} \geq 0$ for $\boldsymbol{n} \in S$ and $\boldsymbol{u} \in K_{\boldsymbol{n}}$ and using the fact that $\pi_{\boldsymbol{n}}(f) = \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}}$ for $\boldsymbol{n} \in S$, we have the following linear program to solve:

$$
\begin{aligned}
&\text{maximize} \sum_{\boldsymbol{n}\in S} \sum_{\boldsymbol{u}\in K_{\boldsymbol{n}}} r_{\boldsymbol{n}}^{\boldsymbol{u}} x_{\boldsymbol{n}}^{\boldsymbol{u}} \\
\text{subject to}\quad & \textstyle\sum_{\boldsymbol{p}\in S} \sum_{\boldsymbol{u}\in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}} - \sum_{\boldsymbol{i}\in S} \sum_{\boldsymbol{u}\in K_{\boldsymbol{i}}} q_{(\boldsymbol{i},\boldsymbol{n})}^{\boldsymbol{u}} x_{\boldsymbol{i}}^{\boldsymbol{u}} = 0 \text{ for } \boldsymbol{n} \in S \qquad (10) \\
& \textstyle\sum_{\boldsymbol{n}\in S} \sum_{\boldsymbol{u}\in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} = 1 \\
& x_{\boldsymbol{n}}^{\boldsymbol{u}} \geq 0 \text{ for } \boldsymbol{n} \in S \\
& \textstyle\sum_{\boldsymbol{n}\in S} \sum_{\boldsymbol{u}\in K_{\boldsymbol{n}}} c_{\boldsymbol{n}}^{\boldsymbol{u}}(l,s) x_{\boldsymbol{n}}^{\boldsymbol{u}} \leq \epsilon \text{ for } 1 \leq l \leq M, 1 \leq s \leq M \; l \neq s. \qquad (11)
\end{aligned}
$$

One of the equations among the set of constraints in Equation (10) can be written in terms of the other equations. Therefore, we can drop one of the constraints in Equation (10). The above optimization problem is a linear program and can be solved using well-known techniques, like the simplex algorithm. The simplex algorithm will reveal if the conditions in Equation (11) are too stringent, which implies that the original class-level QoS constraints in Equation (2) are too stringent. Policy iteration and value iteration techniques [17] for solving Markov decision processes cannot be applied to solve this problem due to the additional constraints in Equation (11).

To obtain the optimal $d_{\boldsymbol{n}}^{\boldsymbol{u}*}$ of the problem, we note that $d_{\boldsymbol{n}}^{\boldsymbol{u}} = x_{\boldsymbol{n}}^{\boldsymbol{u}} / \sum_{\boldsymbol{u}\in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}}$. Therefore, once the $x_{\boldsymbol{n}}^{\boldsymbol{u}*}$ have been computed, the optimal control action in each state is obtained according to the following procedure [14]. Define $S^* = \{\boldsymbol{n} \in S : x_{\boldsymbol{n}}^{\boldsymbol{u}*} > 0 \text{ for some } \boldsymbol{u} \in K_{\boldsymbol{n}}\}$. Then the optimal action in state $\boldsymbol{n}$ is determined as follows. For $\boldsymbol{n} \in S^*$, action $\boldsymbol{v}$ is chosen with probability $x_{\boldsymbol{n}}^{\boldsymbol{v}*} / \sum_{\boldsymbol{u}\in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}*}$, and for $\boldsymbol{n} \notin S^*$, an arbitrary action $\boldsymbol{v} \in K_{\boldsymbol{n}}$ is chosen with probability one. Each time the system visits a state $\boldsymbol{n} \in S^*$ that has more than one action with non-zero probability, an action is chosen according to the probability distribution $d_{\boldsymbol{n}}^{\boldsymbol{u}*}$.

It has been shown in [14] that for a Markov decision process with k additional constraints (apart from the ones that a valid steady-state probability distribution has to satisfy), if the above procedure is used to obtain the optimal policy, then there is a list of at most $|S| + k$ actions from which the optimal policy is chosen. This implies that the optimal policy obtained by the above procedure has at most k states that require randomization. The number of additional constraints in our problem is $M(M-1)/2$ because half the constraints in Equation (11) are automatically satisfied if the other half are satisfied. Therefore, the optimal policy obtained by the above procedure requires randomization in at most $M(M-1)/2$ states, and our numerical results agree with this result. Our example in Section 4 also shows that the admission control scheme designed using the above procedure ensures that the deviation from the desired class-level QoS does not exceed the threshold ϵ.

3 Providing Class-level QoS Subject to a Minimum Throughput Constraint

In this section, we develop an admission control algorithm that minimizes the deviation from the desired class-level QoS subject to a constraint on the throughput. The reference scenario is the same as described in the previous section. Our objective is to obtain an admission control policy that is the solution to the following optimization problem:

$$
\begin{aligned}
&\min_{f\in\mathcal{F}'} \max\left\{|w_{ls}B_l(f) - w_{sl}B_s(f)| : 1 \leq l < M, l+1 \leq s \leq M\right\} \qquad (12) \\
&\text{subject to} \quad T(f) \geq T_{min},
\end{aligned}
$$

where $T(f)$ denotes the throughput of policy f, and T_{min} is the minimum throughput that we want the admission control policy to achieve. As in the previous section, we are justified in restricting our search for optimal policies among stationary policies $\mathcal{F}$. We now rewrite the objective in Equation (12) as

$$\min_{f \in \mathcal{F}} \max\{w_{ls}B_l - w_{sl}B_s : 1 \leq l \leq M, 1 \leq s \leq M, l \neq s\}, \tag{13}$$

where we drop the absolute value signs by taking the maximum over $M(M-1)$ quantities.

The above problem can be formulated as a continuous-time Markov decision process with constraints. The state space, action space, and transition rates of the problem are the same as those described in the previous section. We now identify the (per-unit-time) cost for staying in state $\boldsymbol{n} \in S$ under control action $\boldsymbol{u} \in K_{\boldsymbol{n}}$. We define a vector of costs, instead of a single cost, for staying in state $\boldsymbol{n}$ under control action $\boldsymbol{u} \in K_{\boldsymbol{n}}$. We define our vector in such a way that the long-run average of the first component yields $w_{12}B_1 - w_{21}B_2$, the long-run average of the second component yields $w_{13}B_1 - w_{31}B_3$, and so on. Such a cost vector $c_{\boldsymbol{n}}^{\boldsymbol{u}}$ is determined according to the following algorithm.

Algorithm: Determination of cost vector

1. Initialize $i = 1$.
2. For $l = 1, 2, \ldots, M$ do the following:
3. For $s = 1, 2, \ldots, M$ and $s \neq l$ do the following:
4. Set $c_{\boldsymbol{n}}^{\boldsymbol{u}}(i) = c_{\boldsymbol{n}}^{\boldsymbol{u}}(l, s)$ according to Equation (4) and $i = i + 1$.

The cost vector, as can be seen from the above algorithm, has $M(M-1)$ components with one term each for the $M(M-1)$ terms in the objective function in Equation (13). A control action that accepts class-1 calls but rejects class-2 calls contributes $-w_{21}$ to $w_{12}B_1 - w_{21}B_2$, while a control action that accepts both class-1 and class-2 calls does not contribute anything to $w_{12}B_1 - w_{21}B_2$. Using this definition of cost vector, the term $w_{12}B_1 - w_{21}B_2$ can be written as $\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} \pi_{\boldsymbol{n}}(f) c_{\boldsymbol{n}}^{\boldsymbol{u}}(1) d_{\boldsymbol{n}}^{\boldsymbol{u}}$, where the $\pi_{\boldsymbol{n}}(f)$, the $d_{\boldsymbol{n}}^{\boldsymbol{u}}$, etc., are as defined in the previous section. To meet the constraint on the throughput, we use the same reward definition as in Equation (3). The minimum throughput constraint can now be written as $\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} \pi_{\boldsymbol{n}}(f) r_{\boldsymbol{n}}^{\boldsymbol{u}} d_{\boldsymbol{n}}^{\boldsymbol{u}} \geq T_{min}$. The value of T_{min} has to be chosen appropriately to prevent the problem from becoming infeasible. For example, if the value of T_{min} is too high, then even the admission control policy that maximizes throughput (without any constraints) will not achieve T_{min}. In the numerical results, we choose T_{min} to be αT_{cs}, where $\alpha < 1$, and T_{cs} is the throughput achieved by the complete sharing scheme.

We can convert our problem to $M(M-1)$ linear programming problems by successively considering one of the $M(M-1)$ terms in Equation (13) as the maximum among all the $M(M-1)$ terms. The following is one such linear program obtained by setting $x_{\boldsymbol{n}}^{\boldsymbol{u}} = \pi_{\boldsymbol{n}}(f) d_{\boldsymbol{n}}^{\boldsymbol{u}} \geq 0$ for $\boldsymbol{n} \in S$ and $\boldsymbol{u} \in K_{\boldsymbol{n}}$:

$$\begin{aligned}
&\text{minimize} \sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} c_{\boldsymbol{n}}^{\boldsymbol{u}}(1) \\
&\text{subject to} \quad \textstyle\sum_{\boldsymbol{p} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} q_{(\boldsymbol{n},\boldsymbol{p})}^{\boldsymbol{u}} - \sum_{\boldsymbol{i} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{i}}} q_{(\boldsymbol{i},\boldsymbol{n})}^{\boldsymbol{u}} x_{\boldsymbol{i}}^{\boldsymbol{u}} = 0 \text{ for } \boldsymbol{n} \in S \\
&\textstyle\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} = 1 \\
&x_{\boldsymbol{n}}^{\boldsymbol{u}} \geq 0 \text{ for } \boldsymbol{n} \in S \\
&\textstyle\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} r_{\boldsymbol{n}}^{\boldsymbol{u}} \geq T_{min} \\
&\textstyle\sum_{\boldsymbol{n} \in S} \sum_{\boldsymbol{u} \in K_{\boldsymbol{n}}} x_{\boldsymbol{n}}^{\boldsymbol{u}} (c_{\boldsymbol{n}}^{\boldsymbol{u}}(1) - c_{\boldsymbol{n}}^{\boldsymbol{u}}(j)) \geq 0 \text{ for } j = 2, 3, \ldots, M(M-1).
\end{aligned} \tag{14}$$

Essentially, the above problem minimizes the term $w_{12}B_1 - w_{21}B_2$ subject to the condition that it is the largest among all the $M(M-1)$ terms in Equation (13). We solve $M(M-1)$ such linear programs and the solution corresponding to the problem that yields the least objective function value is the optimal x_n^{u*}.

The number of constraints in our problem is $|S| + M(M-1)$. But $M(M-1)/2 - 1$ of the constraints in (14) are automatically satisfied if the others in (14) are satisfied. Therefore, there is a list of at most $|S| + M(M-1)/2 + 1$ actions from which the optimal policy is chosen. Consequently, the optimal policy requires randomization in at most $M(M-1)/2 + 1$ states. Our numerical results agree with this observation.

An important feature of the algorithms described in Sections 2 and 3 is that they do not require real-time computation of the admission control decision. The decisions may be computed beforehand and a table look-up can be done when needed. Besides reducing computational complexity, the table look-up procedure also reduces the delay before an admission decision is made.

4 Numerical Results

In this section, we present numerical results and compare the performance of our admission control algorithms with that of the throughput maximizing and the complete sharing admission control schemes. We choose fair blocking as the desired class-level QoS requirement for our numerical example. We consider the case of two classes of calls with the total capacity of the channel being 25 BWUs. We also assume that class-1 calls require 1 BWU and class-2 calls require 5 BWUs, i.e., $e_1 = 1$ BWU, and $e_2 = 5$ BWUs. As before, calls arrive according to a Poisson process with λ_1 and λ_2 being the call arrival rates of class-1 and class-2 calls, respectively. The call holding durations for both classes of calls are assumed exponential with mean $1/\mu = 60$ seconds. We fix the call arrival rate of class-2 calls at 0.5 calls/second and treat λ_1 as a parameter.

In Figure 1, we plot the throughput obtained for the various schemes. The constraint on the minimax admission control policy (described in Section 3) is such that T_{min} is chosen to be 99% of the throughput obtained with the complete sharing policy. As expected, we find that the scheme that maximizes throughput performs the best. However, the better performance comes at the cost of a large difference in the call blocking probabilities of the two call classes, as seen in Figure 2. We plot the difference in call blocking probabilities for the schemes under identical conditions in Figure 2. We see that our fairness constraint ($|B_2 - B_1| \leq \epsilon$) ensures that the difference between the call blocking probabilities does not exceed the threshold we set with only a moderate sacrifice in the throughput. The values of $|B_2 - B_1|$ for the scheme described in Section 3 (minimax policy) are many orders of magnitude smaller than the scales shown in the figure.

5 Conclusions

In this work, we addressed the problem of providing class-level QoS to classes of users that have differing bandwidth requirements. We formulated two different problems that consider the two important criteria of providing class-level QoS and improving throughput together. One was the problem of maximizing throughput subject to the constraint that the maximum deviation from the desired class-level QoS be smaller than a pre-specified threshold. The other was the problem of minimizing the deviation from the desired class-level QoS relations subject to the constraint that the throughput achieved does not fall below a certain minimum threshold. We converted the first problem into a linear program and the second to $M(M-1)$ linear programs, where M is the number of call classes.

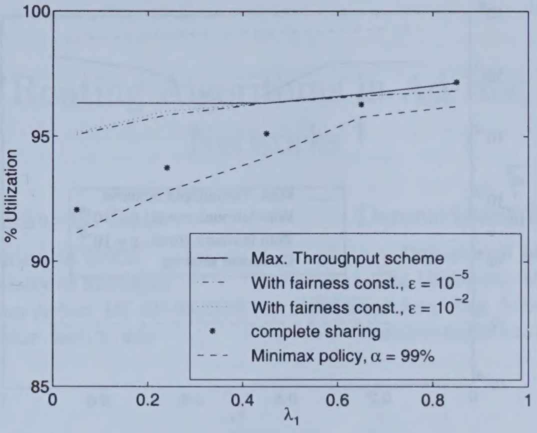


Figure 1: The throughput obtained using various schemes

The numerical results show that we can achieve a significant performance improvement in the class-level QoS provided by the network for only a moderate loss in throughput.

References

[1] J. Liebeherr, "Multimedia networks: Issues and challenges," *Computer*, vol. 28, no. 4, pp. 68–69, April 1995.

[2] D. Mitra, J. A. Morrison, and K. G. Ramakrishnan, "ATM network design and optimization: A multirate loss network framework," *IEEE/ACM Transactions on Networking*, vol. 4, no. 4, pp. 531–543, August 1996.

[3] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1091–1100, August 1995.

[4] E. W. Knightly and N. B. Shroff, "Admission control for statistical QoS: Theory and practice," *IEEE Network*, vol. 13, no. 2, pp. 20–29, March 1999.

[5] G. de Veciana, C. Courcoubetis, and J. Walrand, "Decoupling bandwidths: A decomposition approach to resource management in networks," in *IEEE INFOCOM '94*, vol. 2, (Toronto, Canada), pp. 446–474, 1994.

[6] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, October 1981.

[7] D. Mitra, M. I. Rieman, and J. Wang, "Robust dynamic admission control for unified cell and call QoS in statistical multiplexers," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 692–707, June 1998.

[8] S. C. Borst and D. Mitra, "Virtual partitioning for robust resource sharing: Computational techniques for heterogeneous traffic," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 5, pp. 668–678, June 1998.

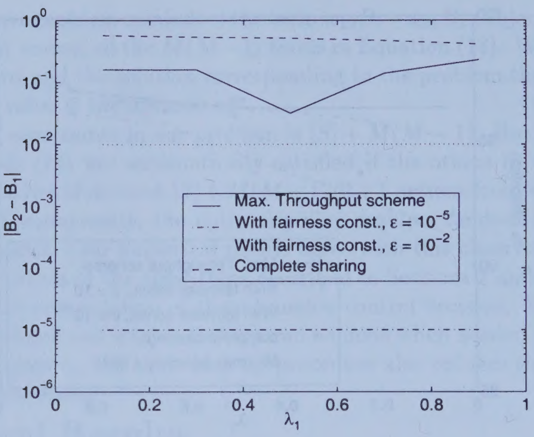


Figure 2: Comparison of $|B_2 - B_1|$ obtained for various schemes

[9] B. Kraimeche and M. Schwartz, "Circuit access control strategies in integrated digital networks," in *IEEE INFOCOM '84*, (San Francisco, CA), pp. 230–235, 1984.

[10] J. M. Hyman, A. A. Lazar, and G. Pacifici, "A separation principle between scheduling and admission control for broadband switching," *IEEE Journal on Selected Areas in Communications*, vol. 11, no. 4, pp. 605–616, May 1993.

[11] Z. Dziong and L. G. Mason, "Fair-efficient call admission control policies for broadband networks – a game theoretic framework," *IEEE/ACM Transactions on Networking*, vol. 4, no. 1, pp. 123–136, February 1996.

[12] R. J. Gibbens and F. P. Kelly, "Resource pricing and the evolution of congestion control," *Automatica*, vol. 35, no. 12, pp. 1969–1985, December 1999.

[13] S. M. Ross, *Applied Probability Models with Optimization Applications.* Holden-Day, 1970.

[14] K. W. Ross, "Randomized and past-dependent policies for Markov decision processes with multiple constraints," *Operations Research*, vol. 37, no. 3, pp. 474–477, May-June 1989.

[15] S. M. Ross, *Introduction to Stochastic Dynamic Programming.* Academic Press, 1983.

[16] H. Mine and S. Osaki, *Markovian Decision Processes.* American Elsevier Publishing Company Inc., 1970.

[17] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley and Sons, 1994.

Optimal Routing Algorithms in Ad Hoc Wireless Networks [†]

Christopher G. Lott
Department of EECS
The University of Michigan
1301 Beal Ave, Ann Arbor, MI 48109-2122
`clott@eecs.umich.edu`

Demosthenis Teneketzis
Department of EECS
The University of Michigan
1301 Beal Ave, Ann Arbor, MI 48109-2122
`teneketzis@eecs.umich.edu`

Abstract

We investigate a network routing problem where a probabilistic local broadcast model for wireless transmission is used. We present results showing that an index policy is optimal for this problem. We extend the original model to allow for power control, and assert that the index nature of the optimal routing policy remains unchanged. Finally, we present two distributed implementations of the routing policy and discuss their convergence properties.

1 Introduction

As communication networks become ever more prevalent, research to improve their design has become both more encompassing and more application specific. Protocol (policy) issues such as service priority, retransmission schemes, flow control, and routing are now studied for large distributed systems and for a variety of communication channel types. Further, it has long been recognized that integrating various source types, such as voice, video, email, web access, and business transaction data into one seamless transparent network results in both the highest efficiency and ease of use. In such networks the ability to provide appropriate Quality of Service (QoS) to each service type is of paramount importance. A further challenge is posed by wireless networks, where these service goals must be achieved with unreliable and time-varying channels, and where new concerns, such as energy consumption and channel interference, impose additional constraints.

The term *ad hoc* is often now applied to networks in which there is no central network controller, each node can itself act as a store-and-forward router, and in which the connectivity topology is time-varying (e.g. see [7]). Such a network is in contrast to, for example, a cellular network where each cell has a central base station through which all data is transmitted. We think of the network as having a number of messages, each of which is destined for some set of destinations. The general routing problem in an ad hoc network is to define a policy which, given the trajectory histories of all the messages, chooses which nodes should next transmit which message. It is also desired that this policy be implementable in a distributed fashion in the network, so the transmission decisions can be decided locally without knowledge of other parts of the network.

[†] This research was supported in part by ARO Grant DAAH04-96-1-0377, AFOSR Grants F49620-96-1-0028 and F49620-98-1-0370, and NSF Grant ECS-9979347.

Many approaches can be taken to network routing optimization. A typical one is maximizing the overall throughput the network achieves. But in many cases for the wireless environment other considerations are at least as important. For example, often in a wireless network the energy source is locally stored in a battery at each node, and the major design goal is to achieve satisfactory communication while using up as little energy as possible. Also, there may be QoS issues, such as delivery timeliness and message priority, which affect the optimal service policy.

This paper briefly explores these design issues in network routing algorithms for ad hoc wireless networks, and provides a novel system model which allows for optimal design in a number of realistic situations. It is organized as follows. In the remainder of Section 1 we briefly discuss the literature available on routing in ad hoc wireless networks, and present a summary of the contributions of our paper. Section 2 contains a qualitative description of our transmission model. In Section 3 we present the basic time-invariant routing problem. In Section 4 we discuss the time-invariant routing problem with power control. In Section 5 we introduce a distributed algorithm for the time-invariant routing problem. In Section 6 we introduce another distributed algorithm for the time-invariant routing problem. Conclusions are discussed in Section 7. A full presentation of the topics discussed in this paper, including literature survey and references, model description, and the detailed analysis of stochastic routing for various models, can be found in [10].

1.1 Ad Hoc Wireless Network Routing Literature

Much research work on routing for ad hoc wireless networks has been published in recent years. Most of this work includes simulation results, comparing the proposed protocol with a generic link-state or distance vector protocol, or comparing specific algorithms against each other (a recent example is [3]). A minority of the published papers include analytic results, such as proofs of convergence or convergence rates. The work on routing for ad hoc wireless networks can perhaps most conveniently be categorized as either on-demand or route maintenance.

- *On-Demand Protocols*
  On-demand protocols are the algorithms which only update routing tables when a new message arrives. Examples include [4], [5], [8], [11], [13], [14]. and [15].
- *Route Maintenance Protocols*
  Route maintenance protocols periodically update topology information in the network independent of traffic flow. Examples include [2], [6], [12], [14]. [16], [17], and [18].

1.2 Contribution

We present what is, to the best of our knowledge, the first network routing protocol which uses a probabilistic local broadcast model for wireless transmission. This model allows for routing decisions to be made based on immediate feedback from actual transmissions. In contrast to most of the existing literature, where link-level communication is point-to-point, we fully utilize the result of each local broadcast and use the information provided by the immediate feedback to construct an optimal routing protocol. Such a protocol is well-suited for a distributed implementation, as it is based only on information local to a node. We present a method describing how the protocol is updated when topological changes occur in a distributed mobile network, and this demonstrates the protocol's practicality.

We extend the model to allow for transmission power control at each node, and show how the fundamental nature of this optimal protocol doesn't change. For this case, an optimal protocol effectively resolves the tradeoffs between fewer long hops vs. more short hops.

We also analyze a time-varying network, where reward for message delivery depends on time, and present conditions under which we can specify an optimal protocol. This model captures the important tradeoff between QoS requirements, such as delivery timeliness, and other criteria, such as energy efficiency.

2 Model Description

We consider the transmission of one message from a source to a destination. As is standard in many mobile radio environments, we assume each mobile has an omnidirectional antenna. Multiple neighboring nodes can in parallel receive a single transmission. We assume neighboring nodes can exchange control messages at minimal cost, implementing the transmission protocol.

The communication channel is degraded by numerous effects, such as blockage, shadowing, sky and receiver noise, and signal interference. Upon transmission failure, there is a rate of message retransmission, which is a function of message length, transmission rate, and protocol.

In our model we assume that the channel degradations can be lumped into just two categories: 1) "Fast" effects. Given what the controller knows, these effects are independent from one transmission to the next. Examples include shadowing, sky and receiver noise, and interference from other mobiles (assumed unpredictable). 2) "Slow" effects. These remain fixed over the entire time of transmission of a message to its final destination. This includes blockage and signal interference with long-term presence in the environment.

We incorporate these channel degradation effects into a probabilistic transmission model. The above assumptions then mean that transmission events are independent from one time to the next. However, reception at separate nodes can be correlated. These points are made explicit in the model described in the next section.

3 The Stochastic Routing Problem

We begin by briefly defining notation and definitions for the system model under consideration, which we refer to as Model (**M**).

N is the number of nodes in the network.

$\Omega = \{1, \ldots, N\}$, the set of all nodes. So $|\Omega| = N$.

$S \subseteq \Omega$ refers to a state of the system, defined as the set of nodes which have received the message.

$R : 2^{\Omega} \rightarrow \Re$ is the reward function, and $R_i := R(\{i\})$.

π is a Markov policy. We write $\pi(S) = i$ to indicate policy π transmits at node i when in state S. We write $\pi(S) = r$ to indicate policy π retires and receives reward $R(S)$ when in state S.

$V^\pi(S)$ is the expected reward when starting in state S under policy π. We often write V_i^π for $V^\pi(\{i\})$.

We write $P^i(S'|S)$ to indicate the probability of reaching state S' from state S when choosing i for transmission, $i \in S$.

We let $P_{ij} := \sum_{S:i,j \in S} P^i(S|i)$. j is called a *neighbor* of i if $P_{ij} > 0$. $\mathcal{N}(i)$ is the set of all neighbors of i, together with i itself.

Definition 3.1 *A function $f : 2^\Omega \to \Re$ is an* index function *on Ω if f satisfies*

$$f(S) = \max_{i \in S} f(\{i\}) \quad \forall S \subseteq \Omega \tag{1}$$

We begin with a centralized version of the stochastic routing problem, formulated as follows.

Problem $(\mathbf{P_1})$

We consider the transmission of a single message, from a given initial state S_o (i.e. a given set of nodes) to a set of destination states, in a wireless ad hoc network of N nodes described by Model **(M)**. Transmission instances occur at discrete time points. Each transmission from a given node i incurs a fixed energy cost c_i. At each transmission instance the transmitting node is chosen by a central controller that always knows the current state of the system (i.e. the set of nodes that have the message). Node transmissions are local broadcasts, that is, multiple neighbor nodes may all simultaneously receive the message. Given the node chosen to transmit, the probability that a given set of nodes receives the message is known and fixed. The central controller is informed, without any cost, as to which nodes receive the message. Nodes never forget the message after they receive it until the end of the transmission process. Each transmission event is assumed independent of those before and after. Control information flow between the nodes and the controller is considered free in energy and instantaneous in time. A reward function R is specified, where R must be an index function (see Def. 3.1). At any instance, the central controller can terminate the transmission process or choose to continue transmitting. The objective is to choose: (i) the node to transmit at each transmission instance, and (ii) the instance to terminate the transmission process, to maximize

$$E\left\{R(S_f) - \sum_{t=1}^{\tau} c_{i(t)}\right\} \tag{2}$$

where τ is the time when the transmission process is terminated, S_f is the state at τ, and $i(t)$ is the node chosen by the transmission policy at time t.

3.1 Analysis of Problem $(\mathbf{P_1})$

For the analysis of Problem $(\mathbf{P_1})$, we can restrict attention to Markov policies without any loss of optimality. We need the following definitions.

Definition 3.2 *A Markov policy π is a* priority *policy if there is a strict priority ordering of the nodes s.t. $\forall i$ we have $\pi(S \cup \{i\}) = \pi(\{i\}) = i$ or $r, \forall S \subseteq \Omega_i$, where Ω_i is the set of nodes of priority lower than i.*

Definition 3.3 *A priority policy π is called an* index *policy if V^π is an index function on Ω.*

The following theorem is the main structural result of this section. It states that we need only consider index policies for Problem $(\mathbf{P_1})$.

Theorem 3.1 (Index Policy) *[10] There is an optimal Markov policy π for Problem $(\mathbf{P_1})$ which is an index policy.*

We present an algorithm which computes the optimal index policy.

Algorithm 1 *(A Dijkstra-Type Algorithm for an Index Policy)*
Define the sets $\mathcal{A}$ and $\mathcal{X}$ as follows.

Initially: *$\mathcal{A}$ contains the nodes of highest reward in arbitrary order (there must be at least one such node); the action taken by the optimal index policy π on these nodes is r. $\mathcal{X}$ is the unordered complement (w.r.t. Ω) of the set of nodes of highest reward.*

During the construction of optimal policy π: *$\mathcal{A}$ contains a priority list of a set S of nodes, $S \subset \Omega$, together with the action specified by π on each node in S. $\mathcal{X}$ is the unordered complement (w.r.t. Ω) of $\mathcal{A}$.*

The algorithm proceeds as follows.

1. *For each $i \in \mathcal{X}$, compute V_i^π from*

$$V_i^\pi = \max\left\{\frac{-c_i + \sum_{S\supset\{i\}:\pi(S)\neq i} P^i(S|i)V^\pi_{\pi(S)}}{\sum_{S\supset\{i\}:\pi(S)\neq i} P^i(S|i)}, R_i\right\} \tag{3}$$

   assuming that i is the node of next highest index in π.
2. *Choose $i \in \mathcal{X}$ with the highest value of V_i^π, with ties broken arbitrarily. Append this node to the list $\mathcal{A}$ as the next priority node, together with the action specified by (3). Remove i from $\mathcal{X}$.*
3. *If $\mathcal{X}$ is empty, stop. If not, goto step 1.*

In Step 1 the right-hand-side of (3) computes the best expected reward for node i, assuming i is the node of next highest index in π. This computation is feasible because π is a priority policy.

The above algorithm also resembles Klimov's algorithm [9] and has the following feature.

Theorem 3.2 *[10] For Problem (P), Algorithm 1 produces an optimal index policy.*

Note that the computational complexity of Algorithm 1 is $O(N^2)$.

4 The Transmission Control Problem

We define the following variant to Problem $(\mathbf{P_1})$.

Problem $(\mathbf{P_2})$

We consider Problem $(\mathbf{P_1})$, with the following addition. At each time step the central controller chooses a node for transmission, from among the nodes with the message, and a transmission *type* from among the allowable types for that node. To each node and transmission type is associated a transmission cost and a probability that a given set of nodes receives the message. At any time, the controller can decide to stop transmitting and receive the reward for its current state, or continue the transmission process. We seek a policy which maximizes (2) under the conditions of Problem $(\mathbf{P_1})$ and the above addition.

By allowing various transmission types at each node, we allow for a choice among transmission power levels, coding/modulation schemes, antenna directionality, or destination addressing.

4.1 Analysis of Problem $(\mathbf{P_2})$

The fundamental nature of Problem $(\mathbf{P_1})$ is not altered by adding in a choice of transmission types. By appropriate definition of the state-space (each transmission type becomes a node), we can convert Problem $(\mathbf{P_2})$ into one that is equivalent to Problem $(\mathbf{P_1})$. Thus, the nature of the optimal policy of Problem $(\mathbf{P_1})$ is maintained. The details of the specification of the appropriate state-space for Problem $(\mathbf{P_2})$ and its equivalence to Problem $(\mathbf{P_1})$ can be found in [10].

5 A Distributed Algorithm for Problem $(\mathbf{P_1})$

We present an algorithm which computes the optimal solution for Problem $(\mathbf{P_1})$, but which has the further characteristic that computations at each node use only information directly from neighbor nodes. This property is critical to the distributed implementation of the optimal policy in an ad hoc wireless network. We claim convergence of the algorithm to the optimal node ordering and value function under the following constraints.

1. Each node i keeps a current estimate, denoted by V_i, of its own optimal expected reward value. Each node i also stores the best estimate (i.e. most recently transmitted) of each neighbor value, denoted $V_{i,j}$, where $j \in \mathcal{N}(i)$.
2. Information transfer among neighboring nodes consists only of the current V_i value of the transmitting node.
3. Each node's V_i information is transmitted asynchronously.
4. A node's V_i update, defined below, is also asynchronous.
5. It is assumed that each node has knowledge of its own $P(S|i)$ update structure. For example, i estimates $\hat{P}(S|i)$ based on all its communications, both control signals and messages.
6. The energy required to run the algorithm is not included in finding the optimal solution for Problem $(\mathbf{P_1})$.

Specifically, we define the following

Algorithm 2 *At each event time, any number of the following two events can occur.*

Event 1 *A node i receives V_j from a neighbor j, $j \in \mathcal{N}(i)$.*

Event 2 *A node i recomputes V_i using the current $V_{i,j}$ values, as follows.*

$$V_i^n = \max\left\{\max_{\tilde{\pi}}\left\{\frac{-c_i + \sum_{\mathcal{N}(i)\supseteq S\supset\{i\}:\tilde{\pi}(S)\neq i} P^i(S|i)V^n_{i,\tilde{\pi}(S)}}{\sum_{\mathcal{N}(i)\supseteq S\supset\{i\}:\tilde{\pi}(S)\neq i} P^i(S|i)}\right\}, R_i\right\} \tag{4}$$

The maximization in (4) is over all local priority orderings $\tilde{\pi}$ of i and its neighbors. Events 1 and 2 occur infinitely often at each node i.

In [10], we prove the following. As long as the value functions are distinct for each node, then Algorithm 2 converges to these value functions in finite time. If two nodes have identical value functions, then value function convergence is guaranteed only asymptotically, though an optimal policy is determined in finite time.

The above results apply for any set of valid initial V_i estimates, and hence convergence is achieved in diverse situations, such as after abrupt changes in network topology. More specifically, when there are abrupt network topological changes, convergence of Algorithm 2 occurs under the following condition: the topological changes are a transient phenomenon, and message transmissions, receptions, as well as value function updates according to (4) take place during these changes. As a result of this condition, at the end of the changes each node has a set of initial estimates of its value function and of its neighbors' value functions. Initiated with these estimates, Algorithm 2 converges to the correct value functions and an optimal routing policy. This result implies that Algorithm 2 will continue to compute optimal routing polices most of the time even when network topological changes occur continuously, provided that the rate (time-scale) at which these changes occur is much slower than the algorithm's rate of convergence.

Also in [10] we use the specific problem structure to provide an efficient algorithm for the computation in (4), and we relate Algorithm 3 and our method of proof to the distributed Bellman-Ford algorithm [2], and to the more encompassing theoretical work in [1].

We end this section with an example in which, since two nodes have identical value functions, convergence of Algorithm 2 to these correct value functions is only asymptotic.

Example

Consider the system of Figure 1 with parameters $0 < p < 1$, $0 < q \leq 1$, $c_1 = c_2 = 1$, and $R_1 = R_2 = 0, R_3 = R$. That is, Q_3 is the destination node. Assume that $R > \frac{1}{p}$, so the system has a non-trivial optimal policy. Transmission success from either Q_1 or Q_2 to the other two nodes is independent, with probabilities p and q respectively. Assume Algorithm 2 begins with $V_i = V_{ij} = R, \forall i, j$.

As Algorithm 2 runs, the node value computations in Q_1 and Q_2 ping-pong back and forth, as the update of one is transmitted to be used in the other's update. Letting V_1^n denote the update value at Q_1 at the nth such update (i.e. the nth update where V_1 actually changes value), it can be easily shown that

$$\begin{aligned} V_1^n &= \frac{(1-p)^2 q^2}{(p+(1-p)q)^2} \quad V_1^{n-1} + \frac{(pR-1)(p+2(1-p)q)}{(p+(1-p)q)^2} \qquad n \geq 1 \\ &:= A \quad V_1^{n-1} + B \end{aligned} \tag{5}$$

where $V_1^0 = R$, and with A and B as defined by (5). Since $A < 1$, standard difference equation methods [19] yield the closed form solution

$$V_1^n = \left(R - \frac{B}{1-A}\right) A^n + \frac{B}{1-A} \tag{6}$$

Of course, a similar equation holds for V_2^n. We compute $\frac{B}{1-A} = R - \frac{1}{p}$. Since $A < 1$, we see that $\lim_{n \to \infty} V_1^n = R - \frac{1}{p}$, which is the correct value function. But since $A > 0$, this value is not reached in finite time.

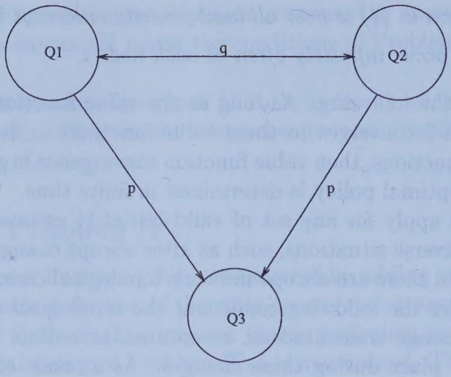


Figure 1: System which takes infinite time to converge

6 A Second Distributed Algorithm for Problem $(\mathbf{P_1})$

We define a second distributed algorithm for Problem $(\mathbf{P_1})$, which is the same as Algorithm 2, except that a different method is used to update the node value function. We define the following

Algorithm 3 *At each event time, any number of the following two events can occur.*

Event 1 *A node i receives V_j from a neighbor j, $j \in \mathcal{N}(i)$.*

Event 2 *A node i recomputes V_i using the current $V_{i,j}$ values, as follows.*

(a) The set of $V_{i,j}$ and V_i values are ranked, high-to-low. A local policy $\tilde{\pi}$ for i is created which uses this ranking.

- *If $V_i \geq V_{i,j}\ \forall j \in \mathcal{N}(i)$, then $\tilde{\pi}$ uses the neighbor ranking, but i is given the* lowest *ranking of all the neighbors (a "reset").*

(b) The following is computed.

$$V_i^n = \max\left\{ \frac{-c_i + \sum_{\mathcal{N}(i)\supseteq S\supset\{i\}:\tilde{\pi}(S)\neq i} P^i(S|i)\, V^n_{i,\tilde{\pi}(S)}}{\sum_{\mathcal{N}(i)\supseteq S\supset\{i\}:\tilde{\pi}(S)\neq i} P^i(S|i)}, R_i \right\} \tag{7}$$

Events 1 and 2 occur infinitely often at each node i.

When i is made the lowest ranked node in $\tilde{\pi}$ in Step (a), and the left-hand term using $\tilde{\pi}$ is the larger term in (7) in Step (b), we call this action a *node reset.*

We conjecture that we are able to prove similar results for Algorithm 3 as we are for Algorithm 2. That is, as long as the value functions are distinct for each node, then Algorithm 3 converges to these value functions in finite time. If two nodes have identical value functions, then value function convergence is guaranteed only asymptotically, though an optimal policy is determined in finite time. Our proofs of these statements are not yet complete, but we expect to finish them shortly.

We believe that there are two potential advantages to Algorithm 3 over Algorithm 2.

1. The computation in (7) is much simpler than that of (4), as maximization over policies is not necessary.
2. The convergence properties of the rank algorithm are quite different. It is a topic for further study how the two algorithms compare in their convergence properties.]

We note the relation of Algorithm 3 to [5]. Major differences include the continuous nature of our updates, and that our node values are interpreted as the expected reward at a given node, whose estimates are constantly being updated by (7). The direction flipping action of [5] corresponds to resetting in Algorithm 3. Proof of convergence of Algorithm 3 is thus made more difficult by the fact that node values can change not just when being reset but also when normally computing updated estimates.

7 Conclusion

We formulated a network routing problem which uses a probabilistic local broadcast model for wireless transmission. We showed that an index policy is optimal for this problem. We extended the model to allow for power control, and showed that the index nature of the optimal policy remains unchanged. Finally, we presented two distributed implementations of the optimal routing policy for the original problem, and discussed their convergence properties. A more detailed version of this work is available in [10], where further results are presented and the claims made in this paper are proved.

Future work in this area can take many forms. For the distributed algorithms, important questions remain, such as convergence rate, counting to infinity, and packet looping. Distance vector algorithms have been in a long process of evolution to deal with these problems, and some of these ideas can perhaps also be used here. Also, methods to estimate the local transmission probabilities at each node should be studied. Finally, a study of the multicast problem for the model of this paper is of great interest, and could lead to more interesting optimization formulations.

References

[1] D. Bertsekas, "Distributed Dynamic Programming", *IEEE Transactions on Automatic Control* **27**:3, p.610-616, June 1982

[2] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1992

[3] J. Broch, D. Maltz, D. B. Johnson, Y. C. Hu, J. Jetcheva, "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols", *Proceedings of the Fourth Annual ACM/IEEE International Conference on Mobile Computing and Networking*, October 25-30, 1998

[4] M. Corson and A. Ephremides, "A Distributed Routing Algorithm for Mobile Wireless Networks", *Wireless Networks* **1**, p. 61-81, 1995

[5] E. Gafni and D. Bertsekas "Distributed Algorithms for Generating Loop-Free Routes in Networks with Frequently Changing Topology", *IEEE Transactions on Communications* **29**:1, p.11-18, 1981

[6] J. Garcia-Luna-Aceves and S. Murthy "A Path-Finding Algorithm for Loop-Free Routing", *IEEE/ACM Transactions on Networking*, v 5 n 1, p.148-160, Feb 1997

[7] Z. HAAS "Guest Editorial, Wireless Ad Hoc Networks", *IEEE Journal on Selected Areas in Communications* **17**:8, p.1329-1330, August 1999

[8] D. B. JOHNSON, "Routing in Ad Hoc Networks of Mobile Hosts", *Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications*, p.158-163, December 1994

[9] G. KLIMOV, "Time Sharing Service Systems I", *Theory of Probability and its Applications*, **19**, p.532-551

[10] C. LOTT AND D. TENEKETZIS, "Stochastic Routing in Ad Hoc Wireless Networks", University of Michigan Control Group Report CGR 00-02 (in preparation)

[11] D. MALTZ, J. BROCH, J. JETCHEVA, D. JOHNSON "The Effects of On-Demand Behavior in Routing Protocols for Multihop Wireless Ad Hoc Networks", *IEEE Journal on Selected Areas in Communications* **17**:8, p.1439-1453, August 1999

[12] J. MOY, "OSPF Version 2", RFC 1247 July, 1991

[13] V. PARK AND M. CORSON, "A Highly Adaptive Distributed Routing Algorithm for Mobile Wireless Networks", *Proceedings of INFOCOM '97*, p.1405-1413, April 1997

[14] M. PEARLMAN AND Z. HAAS "Determining the Optimal Configuration for the Zone Routing Protocol", *IEEE Journal on Selected Areas in Communications* **17**:8, p.1395-1414, August 1999

[15] C. PERKINS "Ad Hoc On Demand Distance Vector Routing", Internet draft, draft-ietf-manet-aodv-00.txt, November 1997. Work in progress.

[16] C. PERKINS AND P. BHAGWAT "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers", *Proceedings of the SIGCOMM '94 Conference on Communications Architectures, Protocols and Applications*, p. 234-244, August 1994

[17] M. PURSLEY, H. RUSSELL, P. STAPLES "Routing for Multimedia Traffic in Wireless Frequency-Hop Communication Networks", *IEEE Journal on Selected Areas in Communications* **17**:5, p.784-792, May 1999

[18] V. RODOPLU AND T. MENG "Minimum Energy Mobile Wireless Networks", *IEEE Journal on Selected Areas in Communications* **17**:8, p.1333-1344, August 1999

[19] H. WILF, *Generatingfunctionology*, 2nd ed, Academic Press, 1994

Analysis of Low Earth Orbit Satellite Networks with Inter-satellite Links

Arunabha Sen*
Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287, USA

Ron Jost
Network Solutions Sector
Motorola Corporation
Chandler, AZ 85248, USA

Jeffrey M. Capone
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287, USA

Abstract

Low Earth Orbit (LEO) satellite networks are increasingly being deployed to deliver communication services for diverse applications (i.e., voice, data and video). The intersatellite network architecture and routing need to be designed to support these diverse quality of service requirements of each application.

In this work, we evalute a network architecture through a recently proposed metric, namely the *Flow Number*. This metric provides a useful tool to the network architect to choose among several competing designs for the crosslink architecture and gives the designer an measure of the minimum crosslink capacity to support communication between all source-destination pairs. A crosslink architecture with a low flow number is more attractive for satellite networks than the one with a high flow number. We show that the flow number of a a hexagonal satellite constellation network with N nodes is $(N/2)^{1.5}/2$.

Once an architecture is chosen, a routing scheme needs to be developed that can support the service needs of each application. To enhance the reliability and congestion control aspects of the network, multiple paths between a source-destination pair must be known. For this purpose, such paths should be either edge or node disjoint. The earlier studies in this area focussed on finding a set of k disjoint (either edge or node) paths between a pair of source-destination nodes such that the sum of these k path lengths is minimized. However, if the path lengths represent the delay on that path, minimizing the sum of the path lengths may not be very meaningful. In this case, we would like to find a set of k disjoint paths between the source-destination pair such that the length of the *longest* path is the *shortest* among all such sets. In this paper we show that finding a set of k edge-disjoint paths between a source-destination pair such that the length of the *longest* path is the *shortest* among all such sets is NP-complete. We give a mathematical programming formulation for the exact solution of the problem. In addition, we give an approximate solution for the problem with a *guaranteed* performance bound.

---

*Corresponding author: Arunabha Sen, Telephone: 480-965-6153, Fax: 480-965-2751, e-mail: asen@asu.edu

Keywords: LEO satellite, intersatellite link, disjoint paths, flow number.

1 Introduction

Low Earth Orbit (LEO) satellite networks are increasingly being deployed to carry out communication of diverse types of traffic [15]. Such networks comprise of ground stations, satellites, ground to satellite links and intersatellite links. Such networks offer a new set of interesting and challenging problems [18]. ATM concepts for such satellite based networks have been proposed in [19].

In [19], the authors introduce the concept of *dynamic virtual topology routing* for routing through networks with time-varying topolgies. In that work, the dynamic network topology is considered to be a periodically repeating series of snapshots of the network topology. Suppose M such snapshots of the topology is considered at intervals of $\Delta t = T/M$, where T is the overall time period. Each of the snapshots at $t = m\Delta t, 0 \leq m \leq M-1$, is modeled as a graph $G(m) = (V, E(m))$, where V is the set of nodes representing the satellites and $E(m)$ represents the set of undirected links between neig'hboring nodes (satellites) existing at time $t = m\Delta t$. Throughout the time period T, a set of unordered $|V|(|V|-1)/2$ source-destination pairs are maintained. The authors suggest assigning a set of $P_w(m)$ of upto k distinct loopless paths to each source-destination pair. Every path $p(m) \in P_w(m)$ comprises of a unique sequence of links $(i,j)_m$. The assignment procedure (whether the link $(i,j)_m$ is part of the path $p(m)$) reflects the set up of an *instantaneous virtual topology* $VT(m)$ within $G(m)$. This task is performed by an iterative k best path search algorithm for every source-destination pair. The authors suggest an iterative approach based on successive calls to the Shortest Path Algorithm due to Dijkstra. They also suggest that by "eliminating" already occupied links, one can force a set of disjoint paths, thus providing a base for simple and robust fault recovery and a framework for traffic flow shaping at the time of operation.

When only considering the intersatellite part of the network, the network topology remains unchanged over time. In this work, two important problems are addresses for routing through the intersatellite part of the network. In the first problem studied in this work, we use a recently proposed graph based metric, the *Flow Number*, to evaluate crosslink architecture of a satellite network. The flow number may be viewed as a measure of the minimum link capacity so that communication between all source-destination pairs can be achieved. This metric is very useful when designing satellite networks and supplements the currently used metrics such as the *diameter* and the *connectivity* for network topology evaluation. A topology with a low flow number is more attractive for satellite networks than one with a high flow number. We show that the flow number of a hexagonal satellite constellation formed with N nodes is $(N/2)^{1.5}/2$.

In the second problem we study disjoint paths in networks. As indicated in the discussion in the earlier, the knowledge of a set of k disjoint (either node or edge) paths between a pair of source-destination nodes can be utilized to enhance the reliability and congestion control aspects of a network. Most of the earlier studies on this topic [1, 5, 9, 11, 14, 16, 17] finds k node or edge disjoint paths between a source-destination node pair such that the sum of the path lengths is minimized. However, if the path lengths represent the delay on that path, minimizing the sum of the path lengths may not be very meaningful. In this case, we would like to find a set of k (node or edge)-disjoint

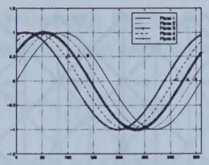


Figure 1: Six planes of LEO constellation illustrating planar relationship

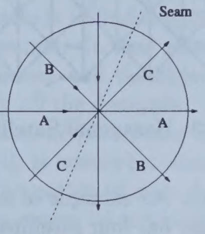


Figure 2: Polar view showing seam

paths between the source-destination node pair such that the length of the *longest* path is *shortest* among all such sets. In this paper we show that finding a set k edge-disjoint paths between the source-destination node pair in an arbitrary graph, such that the length of the *longest* path is *shortest* among all such sets is NP-complete. We give a mathematical programming formulation for the exact solution of the problem. In addition, we give an approximate solution for the problem with a *guaranteed* performance bound.

2 Preliminaries

In a LEO satellite network, each satellite's relationship to their neighbor satellites change as they move through the orbit. The left and right relative satellite neighbors are switched each 180 degrees of the orbit as shown in figure 1. In a satellite constellation there must exist two adjacent planar orbits with satellites moving in opposite orbital directions as shown in figure 2 and this region of the constellation is referred to as the *seam.* Since satellites on either side of the seam are moving in opposite orbital directions, intersatellite communication across the seam is virtually impossible.

If the satellites in the constellation cannot communicate across the seam, then even though the satellite neighbors geographic reference changes (i.e., from right to left) their topological neighbors remain unchanged. Referring to figures 1 and 2, as node A approaches the pole, its right neighbor is node C and its left neighbor is node B. After node A crosses the pole, node B is now its right neighbor and node C is its left neighbor. However, nodes B and C are still neighbors of A. As a result, when only the intersatellite links are considered, the satellite network topology remains unchanged over time. Therefore, routing through the intersatellite part of the network is not a time-vary problem.

A graph where each satellite represents a node and two nodes are adjacent to each other if the satellites represented by these two nodes are connected by an intersatellite link will be referred to as a *satellite graph.* In this paper we consider a satellite constellation where each satellite is connected to six other satellites through intersatellite links. Two of these satellites are in the same satellite plane, two are on the left adjacent plane and

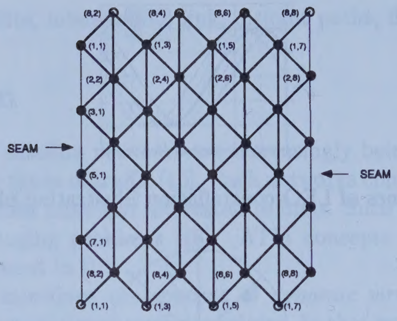


Figure 3: Hexagonal Satellite Grid

two are on the right adjacent plane. An example of such a graph is shown in figure 3.

In the figure, each satellite plane has four satellites indicated by the black nodes and there are eight such planes. In may be noted that the topmost and the bottommost node in one plane is one and the same node. They are drawn this way to indicate that they form an orbit. It may also be noted that every satellite in the constellation has six intersatellite links except the ones on the leftmost and rightmost satellite planes. Since there is no communication across the seam, the satellites on these planes have only four intersatellite links.

3 Flow Number

The metrics like the *connectivity* and the *diameter* of a graph $G = (V, E)$ are used to evaluate the relative merits of different topologies. However, these metrics are unable to capture one important aspect of a network topology. The flow number may be viewed as a measure of the minimum link capacity so that communication between all source destination-pairs can be achieved for a given graph G. This metric is very useful when designing satellite networks and supplements the currently used metrics such as the *diameter* and the *connectivity*, which may be viewed as an indication of the delay through a network topology.

Suppose $d_{i,j}$ is the demand for data stream from node i to node j, $(1 \leq i, j \leq n)$. For example, $d_{i,j}$ may be 1Mbps, implying that 1M bits has to flow from node i to node j per second. Suppose that a demand vector $D = [d_{1,1}, \ldots, d_{n,n}]$ is specified $(d_{i,i} = 0\ \forall i, 1 \leq i \leq n)$. The traffic flow from node i to node j may be routed in many different ways. However, no matter how it is routed, such a flow may be impossible if the capacity (bandwidth) associated with the links is below a certain threshold value. Assuming that the capacity of all the links are identical, we would like to find this minimum threshold capacity of the links such that the demand D can be met through some routing. We refer to this minimum threshold capacity on the links that is able to sustain a traffic flow $d_{i,j} = 1\ \forall i, j,\ i \neq j$, as the *flow number* of the graph $G = (V, E)$.

Formally, the flow number $fn(G)$ of a graph $G = (V, E), (|V| = n, |E| = m)$ may be defined as follows: Suppose that $R_1, \ldots, R_k$ is the set of all possible ways of routing the traffic from the set of sources to the set of destinations. The traffic flow is $d_{i,j} = 1\ \forall i, j,\ i \neq j$. We refer to the traffic from a source i to a destination j as a data stream or a *flow*. Consider a $m \times k$ traffic matix T whose (i, j)-th entry, $T[i, j]$, indicates the number of data streams (or flows) through the edge (link) e_i of the graph $G = (V, E)$ under the routing scheme R_j. Then the flow number of $G = (V, E)$ is

$$fn(G) = \min_{1 \leq j \leq k} [\max_{1 \leq i \leq m} T[i, j]]$$

It may be noted that the flow number of a graph can be computed using the techniques known for the *multicommodity flow* problem [7]. In a sense, it can be viewed as the dual of the multicommodity flow problem. In the multicommodity flow problem, the capacity of the links is known and the objective is to find the maximum flow that can be attained, subject to the link capacity constraints. In flow number computation, traffic demand for all source-destination pair is known and the objective is to find the minimum link capacity that will be able to satisfy all the demands.

It may be noted that the notion of *forwarding index* of communication networks introduced by Chung *et. al.* in [4], is somewhat similar to the notion of *flow number* introduced by Sen *et.al.* in [12]. The notion of *node flow number* of a graph $G = (V, E)$, introduced here is the same as the notion of *minimum achievable forwarding index for networks restricted to G* of [4].

It is well known that the multicommodity flow problems are NP-complete [6]. Approximation algorithms for these problems with a guaranteed performance bound is presented in [8, 13]. In the following two subsections, we show that if the graph has a regular structure, the computation of the flow number of a graph is not necessarily all that difficult. We compute the flow number of a satellite constellation where each satellite (except the ones on the orbit boundary) has six intersatelite links. We show that the flow number of a such a constellation with N nodes is $(N/2)^{1.5}/2$.

We compute the flow numbers in the following way: First we compute a lower bound of the flow number and then we compute an upper bound on the flow number and finally we show that these two are equal. Each undirected edge of is replaced by two oppositely directed edges. We compute the minimum capacity on these directed edges that will be able to satisfy the traffic demand of one unit for every source-destination pair ($d_{i,j} = 1 \ \forall i, j, \ i \neq j, d_{i,i} = 0 \ \forall i$).

3.1 Flow Number of Hexagonal Satellite Constellation

Consider the $M \times M$ hexagonal satellite constellation network shown in figure 3 ($M = 8$). It may be noted that the number of satellites in the constellation is $N = 1/2M^2$. Each undirected link of the figure can be thought of as comprising of two oppositely directed links. We show them as undirected links just to keep the diagram uncluttered. For the purpose of analysis we will consider an undirected link as comprising of two oppositely directed links. For simplicity of analysis, we assume that $M = 4m$ for some positive integer m.

Lemma 1 *The lower bound on flow number of a hexagonal satellite constellation network with N nodes is* $(N/2)^{1.5}/2$.

Proof: Omitted due to page restriction.

Lemma 2 *The upper bound on flow number of a hexagonal satellite constellation network with N nodes is* $(N/2)^{1.5}/2$.

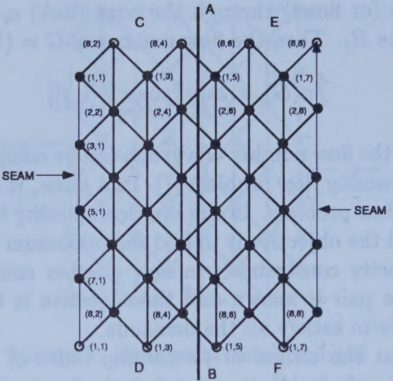


Figure 4: Hexagonal Satellite Grid with the cutline

Proof: The upper bound is established by first describing a routing algorithm. We show that if this algorithm is executed, no directed edge (link) will be required to carry more than a certain number of data streams or flows. This establishes an upper bound on the flow number.

Assume that the satellite grid is divided into two halves by the line AB as shown in figure 4. There are $M^2/4$ nodes on each of these two halves. The $M^2/4$ nodes are distributed in M rows, with each row having $M/4$ nodes. These $M/4$ nodes (the nodes (1,1) and (1,3) in figure 4) have to transfer one data stream to the nodes on the right side (the nodes (1,5) and (1,7) in figure 4). The links $(2,4) \rightarrow (1,5)$ can be used for this purpose. Therefore, the traffic on this links will be $M/4 * M/4 = M^2/16$.

The routing algorithm will route the traffic from a node (i,j) on the left half of the network to a node (k,l) on the right half of the network in the following way ($i < k$ and $j < l$). From (i,j) it will follow the path $(i,j) \rightarrow (i+2,j) \rightarrow (i+4,j) \rightarrow \ldots \rightarrow (k,j)$ or $(i,j) \rightarrow (i+2,j) \rightarrow (i+4,j) \rightarrow \ldots \rightarrow (k+1,j)$ depending on whether or not the node (k,j) exists (e.g., the node (3,2) does not exist). ¿From the node (k,j) (or $(k+1,j)$), it follows the path $(k,j) \rightarrow (k+1,j+1) \rightarrow (k,j+2) \rightarrow \ldots \rightarrow (k,l)$ or $(k+1,j) \rightarrow (k,j+1) \rightarrow (k+1,j+2) \rightarrow \ldots \rightarrow (k,l)$.

In such a traffic routing, each link across the cut line AB will carry exactly $M^2/4 * M/4 = M^3/16$ flows. No link in the network will carry more flows than these links. Thus $M^3/16 = (N/2)^{1.5}/2$ is an upper bound on the flow number of the hexagonal satellite constellation network.

Theorem 3 *The flow number of a hexagonal satellite constellation network with N nodes is* $(N/2)^{1.5}/2$.

Proof: Follows from lemmas 1 and 2.

4 Routing in LEO Satellite Networks

To enhance the reliability and congestion control aspects of the network, multiple paths between a source-destination pair must be known. For reliability and congestion control consideration, such paths should be eitheredge or node disjoint. Disjoint path problems, because of their applications in the design of reliable networks, have been extensively studied [1, 5, 9, 11, 14, 16, 17]. An algorithm for computing a set of k *edge-disjoint* paths between a pair of source and destination nodes, with minimum total edge cost is given in [9]. Suurballe in [16] present an algorithm for the same problem for *node-disjoint* paths. An efficient implementation of the edge-disjoint path problem with minimum total edge cost is given in [17]. Skiscim *et. al.* in [14] study the problem of k successively shortest paths problem. However, in this case the paths are not required to be either node or edge disjoint. Nikolopoulos in [11] studies the k node disjoint paths with minimum total edge cost problem by mapping it into a Trellis graph.

Edge Disjoint k-Path Problem - *Optimization Version*

Instance: A graph $G = (V, E)$, with weight $w_{i,j}$ associated with each edge $e_{i,j}$, $(w_{i,j} \geq 0)$, a source node s, a destination node d and an integer k.

Problem: Find a set of k edge-disjoint paths from s to d such that the length of the *longest* path in this set, is the *shortest* among all such sets of k edge-disjoint paths from s to d.

Edge Disjoint k-Path Problem (EDkPP) - *Decision Version*

Instance: A graph $G = (V, E)$, with weight $w_{i,j}$ associated with each edge $e_{i,j}$, $(w_{i,j} \geq 0)$, a source node s, a destination node d and positive integer k and a non-negative number X.

Question: Is there a set of k edge-disjoint paths from s to d such that the length of the *longest* path in this set is less than or equal to X ?

Theorem 4 *The edge disjoint k-path problem is* NP-complete.

Proof: We show EDkPP is NP-complete by considering a restricted version of the problem where $k = 2$. We call the restricted version as the *Edge Disjoint 2-Path Problem (ED2PP)*.

We show ED2PP is NP-complete by giving a transformation from a well known NP-complete problem - the *partition* problem [6].

Proof: It is easy to see that ED2PP $\in$ NP, since a non-deterministic algorithm need only guess two edge disjoint paths P_1 and P_2 and check in polynomial time that the length of the longer path is less than or equal to X.

We will transform PP to ED2PP. Let the set $A = \{a_1, \ldots a_n\}$ and a "size" $s(a_i) \in Z^+$ for each $a_i \in A, 1 \leq i \leq n$, make up an arbitrary instance of the PP. From this instance of the PP, we must construct a graph $G = (V, E)$ with non-negative weights $w_{(i,j)}$, associated with each edge (i, j), source and destination nodes s and d respectively and a non-negative number X, such that the graph $G = (V, E)$ will have two edge disjoint paths from s to d of length at most X, if and only if the elements of the set A can be partitioned in to two groups, so that the sum of the weights in one group is equal to the sum of the weights in the other group.

We construct the graph $G = (V, E)$ in such a way that it has $3n + 1$ nodes,
$(V = \{\{v_i, 0 \leq i \leq n\} \cup \{(u_{i,1}, u_{i,2}),\ 1 \leq i \leq n\}\})$. The directed edges are from
(i) $v_i \to u_{i+1,1},\ 0 \leq i \leq n-1$, (ii) $v_i \to u_{i+1,2},\ 0 \leq i \leq n-1$,
(iii) $u_{i,1} \to v_i,\ 1 \leq i \leq n$, (iv) $u_{i,2} \to v_i,\ 1 \leq i \leq n$.
Such a graph is shown in figure 5. The weights on the edges are as follows:
(i) $w(v_i, u_{i+1,1}) = 0,\ 0 \leq i \leq n-1$, (ii) $w(v_i, u_{i+1,2}) = 0,\ 0 \leq i \leq n-1$,
(iii) $w(u_{i,1}, v_i) = s(a_i),\ 1 \leq i \leq n$, (iv) $w(u_{i,2}, v_i) = 0,\ 1 \leq i \leq n$
The source node is v_0, the destination node is v_n and $X = 1/2 \sum_{a_i \in A} s(a_i)$.

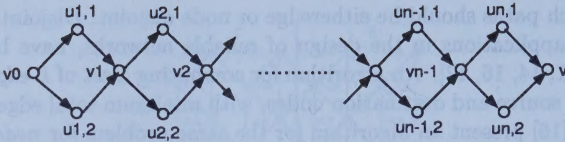


Figure 5: Graph constructed from the instance of the Partition problem

It is easy to see that the construction of the instance of the ED2PP problem can be accomplished in polynomial time. We need to show that instance of the ED2PP problem will have two edge disjoint paths from s to d of length at most X, if and only if elements of the instance of the PP can be divided into two groups, such that the sum of the sizes of the elements in one group is equal to the sum of the size of the elements of the other group.

Suppose that the elements of the set A can be divided into A' and $A - A'$ such that $\sum_{a_i \in A'} s(a_i) = \sum_{a_i \in A-A'} s(a_i)$. In this case we can construct two edge disjoint paths from $s = v_0$ to $d = v_n$ such that the length of the longest path is at most $1/2 \sum_{a_i \in A} s(a_i)$. The paths P_1 and P_2 will have the following directed edges:
(i) $\forall i$, if $a_i \in A'$ the path P_1 will have the edges $v_{i-1} \rightarrow u_{i,1}$ and $u_{i,1} \rightarrow v_i$.
(ii) $\forall i$, if $a_i \in A - A'$ the path P_1 will have the edges $v_{i-1} \rightarrow u_{i,2}$ and $u_{i,2} \rightarrow v_i$.
(iii) $\forall i$, if $a_i \in A'$ the path P_2 will have the edges $v_{i-1} \rightarrow u_{i,2}$ and $u_{i,2} \rightarrow v_i$.
(iv) $\forall i$, if $a_i \in A - A'$ the path P_2 will have the edges $v_{i-1} \rightarrow u_{i,1}$ and $u_{i,1} \rightarrow v_i$.

It can be easily verified that the length of the paths P_1 and P_2 will both be $X = 1/2 \sum_{a_i \in A} s(a_i)$.

Now suppose that the instance ED2PP has two different edge disjoint paths P_1 and P_2 from s to d such that the length of the longer path P_2 is at most $X = 1/2 \sum_{a_i \in A} s(a_i)$. Suppose that the set $N = \{1, 2, \ldots, n\}$. Suppose $M \subseteq N$ is such that the path P_2, is made up of the path segments $v_{m-1} \rightarrow u_{m-1,1} \rightarrow v_m$ for all $m \in M$ and $v_{m-1} \rightarrow u_{m-1,2} \rightarrow v_m$ for all $m \in N - M$. Since the path P_1 is edge disjoint from the path P_2, P_1 is made up of the path segments $v_{m-1} \rightarrow u_{m-1,2} \rightarrow v_m$ for all $m \in M$ and $v_{m-1} \rightarrow u_{m-1,1} \rightarrow v_m$ for all $m \in N - M$. Thus the lengths of the paths P_1 and P_2 are $\sum_{i \in N-M} s(a_i)$ and $\sum_{i \in M} s(a_i)$ respectively. Since the sum of these two path lengths is $\sum_{i \in N} s(a_i) = \sum_{a_i \in A} s(a_i)$ and the length of P_2 is at most $X = 1/2 \sum_{a_i \in A} s(a_i)$, we can conclude that the lengths of the paths P_1 and P_2 are equal to $X = 1/2 \sum_{a_i \in A} s(a_i)$. This implies that $\sum_{i \in M} s(a_i) = \sum_{i \in N-M} s(a_i) = 1/2 \sum_{a_i \in A} s(a_i)$. Thus the elements of the set A can be partitioned into two groups such that the sum of the sizes of the elements of one group is equal to the sum of the sizes of the elements in the other group. This proves the theorem.

Approximate solution to the EDkPP

Consider another edge disjoint path problem with a different objective. This version of the edge disjoint path problem have been studied extensively [9, 16, 17].

Edge Disjoint k-Path Problem - Minimum Total Edge Cost (EDkPP-MTEC)
Instance: A graph $G = (V, E)$, with weight $w_{i,j}$ associated with each edge $e_{i,j}$, $(w_{i,j} \geq 0)$, a source node s, a destination node d and an integer k.
Problem: Find a set of k edge-disjoint paths from s to d such that the sum of the length of the paths in this set, is the *minimum* among all such sets of k edge-disjoint paths from s to d.

Clearly, EDkPP is different from EDkPP-MTEC in its objective. The difficulty of these problems are also different. In this paper, we have shown that the EDkPP is NP-Complete, whereas it is well known that the EDkPP-MTEC can be solved in polynomial time. If n is the number of nodes in the graph and m is the number of edges, then the set of k edge-disjoint paths from a given source s to a given destination d, with minimum total edge cost can be found in $O(km\log_{1+m/n} n)$ time [9, 17].

We use the algorithm to compute the optimal solution to the EDkPP-MTEC problem [9] as an approximate solution to the EDkPP problem. Suppose that the k disjoint paths generated by this algorithm, are $P_1, P_2, \ldots P_k$ and the k disjoint paths that are optimal solution of the EDkPP are $P'_1, P'_2, \ldots P'_k$. We denote the length of the paths P_i and P'_i by PL_i and PL'_i respectively. Without loss of generality, we assume that $PL_i \leq PL_{i+1}$ and $PL'_i \leq PL'_{i+1}$, for $1 \leq i \leq k-1$.

We measure the quality of the approximate solution to the EDkPP by the ratio PL_k/PL'_k.

Since the path set $P_1, P_2, \ldots P_k$ is the optimal solution to the EDkPP-MTEC problem and the path set $P'_1, P'_2, \ldots P'_k$ is the optimal solution to the EDkPP problem,

$$PL_1 + PL_2 + \ldots + PL_k \leq PL'_1 + PL'_2 + \ldots + PL'_k$$

Since $PL'_i \leq PL'_{i+1}$ for $1 \leq i \leq k-1$, the above equation can be rewritten as

$$PL_1 + PL_2 + \ldots + PL_k \leq kPL'_k$$

or $PL_k \leq kPL'_k$ as $PL_i \geq 0$ for $1 \leq i \leq k$. This reduces to

$$PL_k/PL'_k \leq k$$

Thus the ratio of the *approximate solution* to the EDkPP, PL_k, to the *optimal solution*, PL'_k, is bounded by k.

References

[1] R. Bhandari, “Optimal diverse routing in telecommunication fiber networks”, *Proc. IEEE INFOCOM'94*, pp. 1498-1508, 1994.

[2] H. S. Chang em et. al., “Topological design and routing for low earth orbit satellite networks”, *Proc. IEEE Globecom'95*, pp. 529-535, 1995.

[3] S. Chen and K. Nahrstedt, “An overview of QoS routing for next generation high speed networks, problems and solutions”, *IEEE Network*, pp. 64-79, December, 1998.

[4] F. R. K. Chung, E. G. Coffman, M. I. Reiman and B. Simon, “The forwarding index of communication networks”, *IEEE Transactions on Information Theory*, vol. IT 33, no. 2, pp. 224-231, 1987.

[5] D. A. Dunn, W. .D. Grover, M. H. MacGregor, “Comparison of k-shortest paths and maximum flow routing for network facility restoration”, *IEEE Selected Areas of Communications* , vol. 12, no. 1, 1994.

[6] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman Publishers, San Francisco, 1978.

[7] T. C. Hu, Integer Programming and Network Flows, Addison Wesley Publishing Company, 1970.

[8] P. Klein, S. Plotkin, C. Stein, E. Tardos, "Faster approximation algorithms for the unit capacity concurrent flow problem with applications to routing and finding sparse cuts", *SIAM Journal of Computing*, vol. 23, no. 3, pp. 466-487, 1994.

[9] E. L. Lawler, Combinatorial Optimization: Networks and Matroids. Holt, Rinehart and Winston, New York, 1976.

[10] M. D. Noakes, J. B. Cain, J. W. Nieto and E. L. Althouse, "An adaptive link assignment algorithm for dynamically changing topologies", *IEEE Transactions of Communications*, vol. 41, no. 5, pp. 694-706, 1993.

[11] S. D. Nikolopoulos, A. Pitsilides and D. Tipper, "Addressing network survivability issues by finding the k-best paths through a Trellis graph", *Proc. IEEE INFOCOM'97*, pp. 370-377, 1997.

[12] A. Sen, T. Shah, S. Bandyopadhyay and B.P. Sinha, "On new architectures for WDM networks", To appear in the *Proc. of 37th Annual Allerton Conference on Communication, Control and Computing*, University of Illinois at Urbana-Champaign, September 1999, Extended version available as *Techinal Report*, Department of Computer Science and Engineering, Arizona State University, 1999.

[13] F. Shahrokhi and D.W. Matula, "The maximum concurrent flow problem", *Journal of the Association of Computing Machinery*, vol. 37, no. 2, pp. 318-334, 1990.

[14] C. C. Skiscim and B. L. Golden, "Solving k-shortest and constrained shortest path problems efficiently", *Annals of Operations Research*, vol. 20, pp. 249-282, 1989.

[15] M. A. Sturza, "Architecture of the TELEDESIC satellite system", *Proc. of 4th International Mobile Satellite Conference, IMSC'95*, pp. 212-218, Ottawa, Canada, 1995.

[16] J. W. Suuraballe, "Disjoint paths in a network", *Networks*, vol. 4, pp. 125-145, 1974.

[17] J. W. Suuraballe and R. E. Tarzan, "A quick method for finding the shortest pair of disjoint paths", *Networks*, vol. 14, pp. 325-336, 1984.

[18] M. Werner, A. Jahn, E. Lutz and A. Bottcher, "Analysis of system parameters for LEO/ICO satellite communication networks", *IEEE Journal on Selected Areas of Communications*, vol. 13, no. 2, pp. 371-381, 1995.

[19] M. Werner, "ATM concepts for satellite personal communication networks", *Proc. European Conference NOC'96, Heidelberg, Germany*; ATM, Networks and LANs, Eds. D.W. Faulkner and A.L. Hammer, IOS spress, pp. 247-254, 1996.

[20] M. Werner, C. Delucchi, H. J. Vogel, G. Maral, J. J. De Rider, "ATM based routing in LEO/MEO satellite networks with intersatellite links", *IEEE Journal on Selected Areas of Communications*, vol. 15, no. 1, pp. 69-82, 1997.

Control over networks

Hideaki Ishii   Bruce A. Francis

Department of Electrical and Computer Engineering
University of Toronto, Toronto, Canada M5S 3G4
email: {hideaki, francis}@control.utoronto.ca

Abstract

The recent growth in communication technology has allowed control systems to employ networks, so-called control networks, which replace conventional point-to-point connections, between controllers and sensors/actuators. The aim of this paper is twofold: We first discuss the advantages and the drawbacks in the use of such networks and provide an overview on the communication methods of control networks. Then we summarize our recent results that were motivated by the limited data rate constraint that networks impose.

1 Introduction

In large-scale systems, hierarchical structure of several networks is often employed for flexibility and efficiency of communication among subsystems. At the lowest level of the hierarchy of networks is the control network, or the fieldbus, that connects controllers and sensors/actuators. Due to the cost reduction of interfaces and the growth in the communication rate, the use of control networks has become a more common solution in industry, replacing conventional point-to-point connections [18, 19].

The use of such networks reduces the number of cables, which in turn makes system monitoring and maintenance easier. Moreover, it allows the systems to have more modular and flexible structures. Applications using such networks include robotic systems [11], networks in automobiles [12, 15], and jacking systems for train cars [1].

These advantages are available, however, at the cost of limited bandwidth, as for any digital networks. Special design techniques are required so that in any event the given bandwidth is adequate for proper operation; this is particularly important because signals in control systems have critical real-time requirements.

Bandwidth, or data rate, can be characterized by two aspects: time sequencing and quantization. For many components in a system to use a single common network, signals are transmitted sequentially, while all signals have to be represented by a finite number of bits. Clearly, there is a trade-off between these two aspects: If we use fewer bits in quantization, data size decreases, which allows us, in return, to increase the number of data units in a given time period, and vice versa.

Brockett [2] initiated research on some interesting problems of control over networks where the available data rate cannot be idealized as being infinite. Concerning signal sequencing, he proposed a simple model of control networks and a design method of a central controller connected to sensors and actuators by a network. Because of the limitation on data rate, only one of the sensor/actuator pairs is allowed to communicate with the controller at each discrete instant.

On the other hand, systems whose control is based on quantized measurements certainly have a hybrid nature: They have continuous dynamics accompanied by discrete signals taking values in a countable set. However, in hybrid systems, only a few studies are concerned with data rate issues. Elia and Mitter [6, 7] proposed a stabilization technique for linear systems; a nonuniform quantizer is designed as a result of an optimization problem. Brockett and Liberzon [3] studied the use of a uniform but time-varying quantizer to reduce the bit rate. Investigating from an information theoretic viewpoint, Nair and Evans [17] derived stability results for systems with limited data rate.

In this paper, we first give an overview on the communication methods of control networks and then present a summary of our recent works motivated by the recent interest in systems with limited communication due to the presence of networks. The first work deals with the time sequencing aspect of bandwidth [9]. We propose a design problem of dynamic local controllers which are connected to each other by a network; we characterize stabilizability issues and clarify the advantage of using networks.

The second one is on quantization, and we study a stabilization method for a continuous-time, linear plant via switching control. We employ the switching control strategy based on state partitioning. The state space is partitioned into cells and for each cell one has a corresponding constant control input. To prevent chattering, a switching scheme with *dwell time* is employed.

2 Control networks

For control systems that have many agents, the network resource has to be shared in a systematic way. However, the nature of this kind of communication has several distinguishing characteristics, different from that of data communication. First, most of the traffic between controllers and sensors/actuators has a fixed sampling period and is transmitted continuously. Second, the data size of each message is relatively small, and the number of bits required for each signal is known a priori. Third, since time delay has a serious effect on performance, the real-time requirement of fieldbuses is much more critical than that of data networks.

To meet these characteristics, the communication methods used in fieldbuses are designed as follows [18, 4]: Control networks are digital, employing sequential, time-division multiplexing. In particular, they use a communication type with a *periodic pattern*, which consists of periodic and aperiodic phases (Fig. 1). The periodic phase is of signals which have real-time requirements and are periodically transmitted. This phase is deterministic in that the order of signal transmission is scheduled and fixed prior to operation. Hence, no collision of traffic occurs, and the delay time and the necessary data rate are known.

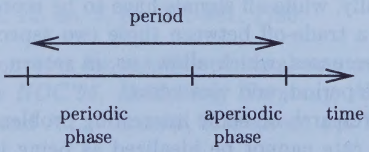


Figure 1: Periodic pattern of communication

The aperiodic phase is composed of signals such as emergency alarms and on/off type signals, which need to be transmitted only occasionally. To keep the system from running into a hazardous state, it is important to design the periodic phase leaving enough transmission time for aperiodic signals so that under any circumstance the delay time of these signals is sufficiently small.

Because of this periodic and deterministic nature of communication and the critical requirement in time, a centralized access method is preferred: a master station controls the traffic on the network and polls agents according to the periodic schedule. As a result, the protocols are often simpler than many of those used for computer networks.

3 Decentralized control problem [9]

In this section, we formulate our design problem for a decentralized control system with two local controllers connected by a network cable. Signals not available locally for each controller can be communicated over the network, possibly at a rate slower than the sampling rate of the controllers. We give a simple model for the network, governed by the protocol described in the previous section, as a "switch box." Because of the sequential transmission over the network, at most one signal is allowed to be communicated through the switch box at each discrete-time instant. The behavior of the switch box is periodic and deterministic.

3.1 The switch box problem

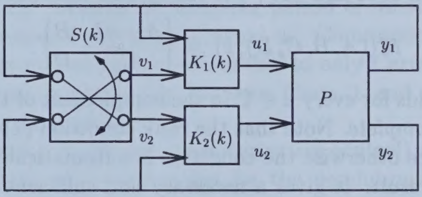


Figure 2: The switch box problem

Consider the discrete-time system in Fig. 2. Plant P is linear time-invariant (TI) with a realization

$$\left\{\begin{array}{rcl} x(k+1) & = & A_p x(k) + B_{p1}u_1(k) + B_{p2}u_2(k), \\ y_i(k) & = & C_{pi}x(k) + D_{pi1}u_1(k) + D_{pi2}u_2(k), \quad i = 1, 2, \end{array}\right.$$

where $x(k) \in \mathbb{R}^n$, $u_i(k) \in \mathbb{R}^{m_i}$, $y_i(k) \in \mathbb{R}^{p_i}$, $k \in \mathbb{Z}_+$. Local controllers $K_i(k)$, $i = 1, 2$ are periodically time-varying (PTV) with period N_K:

$$\left\{\begin{array}{rcl} \xi_i(k+1) & = & A_{ki}(k)\xi_i(k) + B_{ki1}(k)y_i(k) + B_{ki2}(k)v_i(k) \\ u_i(k) & = & C_{ki}(k)\xi_i(k) + D_{ki1}(k)y_i(k) + D_{ki2}(k)v_i(k), \end{array}\right.$$

where $\xi_i(k) \in \mathbb{R}^{n_{ki}}, v_i(k) \in \mathbb{R}^{p_i}$, $k \in \mathbb{Z}_+$; N_K is the smallest positive integer for which $A_{ki}(k) = A_{ki}(N_K + k)$ for all $k \in \mathbb{Z}_+$ and likewise for other system matrices. *Switch box* $S(k)$ is PTV with period N and takes a $(p_1 + p_2) \times (p_1 + p_2)$ matrix of either

$$S_0 := \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad S_1 := \begin{bmatrix} 0 & I_{p_2} \\ 0 & 0 \end{bmatrix}, \quad \text{or} \quad S_2 := \begin{bmatrix} 0 & 0 \\ I_{p_1} & 0 \end{bmatrix}.$$

The switch box provides each controller with the signals not accessible locally. Its periodic behavior is determined by a *switching pattern* $s \in \{0, 1, 2\}^N$, whose elements correspond to the indices of matrices S_i, $i = 0, 1, 2$. For example, $s = [1\ 2\ 0\ 0]$ means $S(k)$ has period $N = 4$ and $S(4k) = S_1$, $S(4k+1) = S_2$, $S(4k+2) = S(4k+3) = S_0$ for $k \in \mathbb{Z}_+$.

We define our *switch box (SB) problem* as follows: Assume that P is stabilizable and detectable. Given a switching pattern $s \in \{0,1,2\}^N$, find PTV local controllers $K_1(k)$ and $K_2(k)$ such that the closed-loop system is exponentially stable.

A technique similar to the switch box is employed in the design problems in [2, 8]. Their setup is that a controller is connected to pairs of sensors and actuators by a network and can communicate with only one pair at each sampling time, depending on a periodic pattern. Even though their model is also simple and linear, because of the constraint on the controller structure (it must be memoryless), the stabilization problem was shown to be an NP hard nonlinear optimization problem. We have extended their problems to a decentralized control setting. Our solution is much simpler because we have enlarged the class of controllers.

3.2 Stabilization using PTV local controllers

Given switching pattern s, our SB problem is a stabilization problem of a PTV plant by local PTV controllers. This problem was studied extensively in [14]; we follow their formulation and give some definitions.

A linear TI system $G = (A, B, C, D)$ is said to be *complete* if

$$\operatorname{rank}(M(A,B,C,D); z) \geq n \quad \text{for } z \in \mathbb{C}, \tag{1}$$

where n is the dimension of A and $M(\cdot;\cdot)$ is the system matrix for G defined by

$$M((A,B,C,D); z) := \begin{bmatrix} A - zI & B \\ C & D \end{bmatrix}.$$

If the rank condition holds for every $z \in \mathbb{C}$ in the complement of the open unit disk, then G is said to be *weakly complete*. Note that the rank condition (1) has to be checked only for eigenvalues of A since otherwise the condition is automatically satisfied.

We state our main result. It gives a necessary and sufficient condition for the solvability of our SB problem under mild conditions.

Theorem 3.1 Suppose that P is stabilizable and detectable,

$$\begin{bmatrix} P_{11} \\ P_{21} \end{bmatrix} \neq 0 \text{ and } \begin{bmatrix} P_{12} \\ P_{22} \end{bmatrix} \neq 0.$$

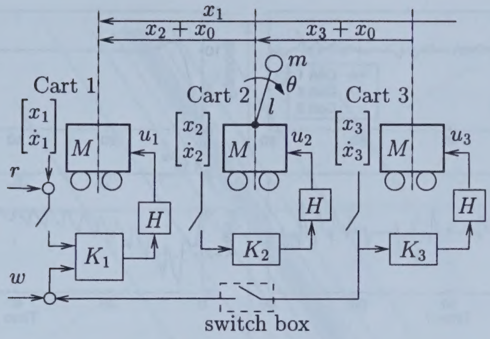
Then P is stabilizable using the network structure in Fig. 2 with PTV local controllers if and only if

(i) $P_{12} = 0$ and is not weakly complete $\Rightarrow$ s contains 1;
(ii) $P_{21} = 0$ and is not weakly complete $\Rightarrow$ s contains 2.

Stabilizing a system in a purely decentralized setup can be difficult (e.g., [21]); however, this result tells us that the transmission of local measurements to other controllers need not be at a high rate if it is only for stabilization. A natural question that arises here is, how does the communication rate affect the capability of the system? This problem is dealt in [9].

3.3 Numerical example

In this section, we give a toy example of autonomous vehicle formation to illustrate our results. This example satisfies the following conditions so that our theory is applicable: The original plant is open-loop unstable, stabilizable, and detectable, and the unstable poles are not assignable without a network structure. So, the system is stabilizable only when signal transmission over the network is allowed.



x_1
$x_2 + x_0$
$x_3 + x_0$
m
θ
l
Cart 1
Cart 2
Cart 3
$\begin{bmatrix} x_1 \\ \dot{x}_1 \end{bmatrix}$
$\begin{bmatrix} x_2 \\ \dot{x}_2 \end{bmatrix}$
$\begin{bmatrix} x_3 \\ \dot{x}_3 \end{bmatrix}$
M
M
M
u_1
u_2
u_3
r
H
H
H
K_1
K_2
K_3
w
switch box

Figure 3: Carts system

Consider the three cart system in Fig. 3. Cart 1 receives reference r for its position and is the leader. Carts 2 and 3 follow it trying to keep the distance x_0 between them; M is the mass of each cart; x_2 and x_3 are the changes in the distance between the carts. Cart 2 has an inverted pendulum of mass m, length l, and angle θ. However, θ is not measurable. Each cart has a discrete-time local controller K_i, $i = 1, 2, 3$, whose input is the measured signal $[x_i, \dot{x}_i]^T$ sampled at sampling period h; its output goes through the 0th order hold H and becomes the control signal u_i. Communication among the carts is to be kept low: signals x_2 and x_3 are measurable to only Carts 2 and 3, respectively; a communications channel is given only between Carts 1 and 3, and its behavior is determined by switching pattern s.

Suppose K_2 is constant (i.e., proportional-derivative control) and is designed so that Cart 2 follows Cart 1 with no consideration for the pendulum. Thus, the pendulum cannot be stabilized locally by Cart 2. The control problem is to design s, K_1, and K_3 so that the whole system, including the inverted pendulum, is stabilized, and the carts follow r. State θ is observable only through signals x_3, $\dot{x}_3$. Hence, its stabilization is possible only by sending signals from Cart 3 to Cart 1 as shown with a switch box in Fig. 3.

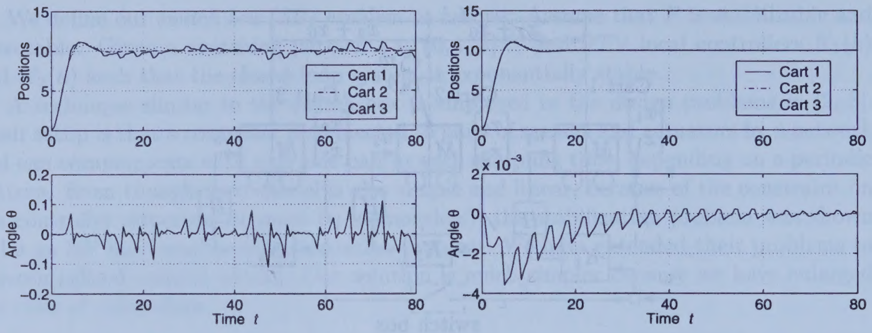
Let v_i be the velocity of Cart i for $i = 1, 2, 3$. Then $v_1 = \dot{x}_1$, $v_2 = \dot{x}_1 - \dot{x}_2$, and $v_3 = \dot{x}_1 - \dot{x}_2 - \dot{x}_3$. The dynamics of the linearized system is described as

$$\begin{aligned} M\dot{v}_1 &= 100u_1, & M\dot{v}_2 &= 100u_2 - mg\theta, \\ M\dot{v}_3 &= 100u_3, & Ml\ddot{\theta} &= (M+m)g\theta - 100u_2, \end{aligned}$$

where $g = 9.8$ m/s^2, $M = 3$ kg, $m = 0.1$ kg, $l = 1$ m, and $K_2 = [0.01\ 0.02]$.

To compare the achievable performance of the systems with different switching pattern, a simple controller was designed using the H^∞ method as follows. Let $K_2 = K_3 = [0.01\ 0.02]$, $N = N_K = 6$, and $h = 0.6$ s. Design a servo controller K_1 so that x_1 follows r, and θ is small with an appropriate weight.

The local controller for Cart 1 was computed for $s_2 = [1\,1\,0\,0\,0\,0]$ and $s_6 = [1\,1\,1\,1\,1\,1]$, where 1 indicates the timing of transmission from Cart 3 to 1. Figure 7 shows step responses of positions of the carts and angle θ for $r(t) = 10$, $t \geq 0$. Here, to emphasize the difference in performance, white noise, uniformly distributed and modeling quantization error, is added to the signals communicated over the network; it is assumed that 23 bits can be used. The system with s_2 is sensitive to noise, showing a noticeable oscillation in cart positions, while the system with s_6 exhibits almost perfect tracking.



(a) For $s_6 = [1\ 1\ 0\ 0\ 0\ 0]$ (b) For $s_6 = [1\ 1\ 1\ 1\ 1\ 1]$

Figure 4: Step responses

4 Quantized control problem [10]

Control strategy based on quantized measurements can be viewed as switching control with state partitioning, as proposed by Delchamps [5]: The state space is partitioned into cells and for each cell one has a corresponding constant control input. We emphasize, as did Elia and Mitter [6], that the cells are not necessarily small, as in the traditional use of the term quantization.

One of the major obstacles in switching control of a continuous-time system is that chattering, or arbitrarily fast switching, may occur. This is not only physically harmful to the actuators and/or the plant, but also impossible to realize with a finite data rate channel. In addition, it presents mathematical difficulties concerning existence of solutions of differential equations.

Of the several methods to prevent chattering [16, 20], we here employ a *dwell-time* switching controller, the protocol being that switching cannot occur until a specified dwell time has elapsed. Some papers treating continuous-time switching controllers, e.g., [3], handle the potential of chattering by permitting solutions in the sense of Filippov [22]. By contrast, our switching with dwell time prevents chattering, and control signals are piecewise constant.

We propose a method to design the partition of the state space of a linear time-invariant, continuous-time plant and the dwell-time switching logic that guarantee stability of the system with a bound on the dwell time. The resulting partition is nonuniform, and its cell size becomes larger logarithmically as the distance from the origin grows. The scheme generalizes Elia and Mitter's [6], which treated the simpler discrete-time case where chattering cannot arise.

4.1 Dwell-time switched systems

In this section, we introduce a class of hybrid systems, called *dwell-time switched systems*, and develop a Lyapunov-like stability result for such systems. We shall use a specific partition, though the stability theory would hold for a more general one. (We don't want to get bogged down in the tricky issue of when to switch were the boundaries of the cells topologically complicated.)

We define a *logarithmic partition* of state space, $\mathbb{R}^n$. Let $e \in \mathbb{R}^n$ be a vector of unit

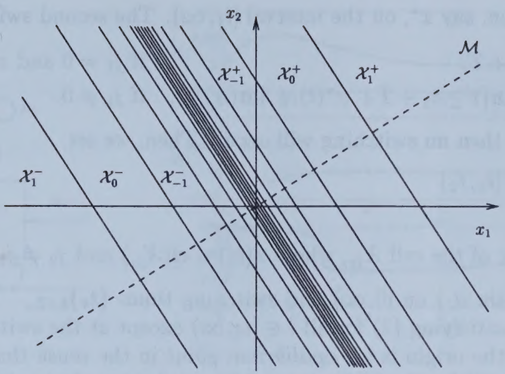


Figure 5: Logarithmic partition of $\mathbb{R}^2$

length, let $\mathcal{M}$ denote the 1-dimensional subspace generated by e, and let $\delta > 1$ and $\alpha_0 > 0$. The partition cells are

$$\mathcal{X}_{00} := \mathcal{M}^\perp \text{ and } \mathcal{X}_j^\pm := \{x \in \mathbb{R}^n : \ \langle x, e\rangle \in \pm[\alpha_0\delta^j, \alpha_0\delta^{j+1})\}, \quad j \in \mathbb{Z}.$$

The picture is Fig. 5 for $n = 2$. Thus, $\mathcal{X}_{00}$ is an $n-1$-dimensional subspace, so it has empty interior, while other cells, being bounded by parallel hyperplanes, have nonempty interiors. For simplicity, we rename the cells to $\mathcal{X}_j, j \in \mathbb{Z}_+$ in the following manner: $\mathcal{X}_0 = \mathcal{X}_{00}$, $\mathcal{X}_1 = \mathcal{X}_1^+$, $\mathcal{X}_2 = \mathcal{X}_1^-$, $\mathcal{X}_3 = \mathcal{X}_{-1}^+, \cdots$.

Now consider the nonlinear switching system

$$\dot{x}(t) = f_{i(t)}(x(t)), \tag{2}$$

where $x(t) \in \mathbb{R}^n$ is the state of the system, and $i(t) \in \mathbb{Z}_+$ is the state of the *switching logic* with a fixed *dwell time* $T > 0$. It is assumed that f_j is globally Lipschitz continuous for each $j \in \mathbb{Z}_+$ and that $f_0(0) = 0$.

The switching logic (how x generates i) is defined in terms of switching times as follows: Set $t_0 = 0$ and let j_0 denote the index of the cell $x(t_0)$ is in. The equation

$$\dot{x}(t) = f_{j_0}(x(t))$$

has a unique solution, say x^*, on the interval $[t_0, \infty)$. If $x(t_0) = 0$, then no switching will ever occur; this is the trivial case. Otherwise, the first switching time is

$$t_1 = \begin{cases} t_0 + T, & \text{if } j_0 = 0 \text{ and } x(t_0) \neq 0 \\ \min\{t \geq t_0 + T : \ x^*(t) \notin \operatorname{int}(\mathcal{X}_{j_0})\}, & \text{if } j_0 \neq 0, \end{cases}$$

where $\operatorname{int}(\cdot)$ is interior of a set. Letting $\operatorname{cl}(\cdot)$ denote closure, we set

$$i(t) = j_0 \text{ for } t \in [t_0, t_1)$$

$$j_1 : \begin{cases} 0 & \text{if } x(t_1) \in \mathcal{X}_0 \\ \text{the index of the cell } \mathcal{X}_{j_1}, \text{ where } x(t_1) \in \operatorname{cl}(\mathcal{X}_{j_1}) \text{ and } j_1 \neq j_0 & \text{otherwise.} \end{cases}$$

This rather complicated definition is to avoid the situation where trajectory x is on the hyperplane $\mathcal{X}_0$ at $t = t_0 + T$ and will leave it the next moment; the cell that the state is entering cannot be defined.

Proceeding, the equation

$$\dot{x}(t) = f_{j_1}(x(t))$$

has a unique solution, say x^*, on the interval $[t_1, \infty)$. The second switching time t_2 is

$$t_2 = \begin{cases} t_1 + T, & \text{if } j_1 = 0 \text{ and } x(t_1) \neq 0 \\ \min\{t \geq t_1 + T : \ x^*(t) \notin \operatorname{int}(\mathcal{X}_{j_1})\}, & \text{if } j_1 \neq 0. \end{cases}$$

Again, if $x(t_1) = 0$, then no switching will occur. Then, we set

$$i(t) = j_1 \text{ for } t \in [t_1, t_2)$$

$$j_2 : \begin{cases} 0 & \text{if } x(t_2) \in \mathcal{X}_0 \\ \text{the index of the cell } \mathcal{X}_{j_2}, \text{ where } x(t_2) \in \operatorname{cl}(\mathcal{X}_{j_2}) \text{ and } j_2 \neq j_1 & \text{otherwise.} \end{cases}$$

Continuing, we obtain $i(\cdot)$ on $[0, \infty)$ and switching times $\{t_k\}_{k \in \mathbb{Z}_+}$. The trajectory $x(\cdot)$ is uniquely defined satisfying (2) for all $t \in [t_0, \infty)$ except at the switching times.

For system (2), the origin is an *equilibrium point* in the sense that if $x(0) = 0$, then $x(t) = 0$ for all $t \geq 0$. It is natural to give stability definitions for such systems as follows.

Definition 4.1 The origin of system (2) is (i) *stable* if for every $\epsilon > 0$ there exists $\delta > 0$ such that if $\|x(0)\| < \delta$ then $\|x(t)\| < \epsilon$ for all $t \geq 0$; (ii) (*globally*) *asymptotically stable* if it is stable and if, for every $x(0) \in \mathbb{R}^n$, $x(t) \to 0$ as $t \to \infty$.

A simple extension of standard Lyapunov stability theory (e.g., [13], Theorem 3.1) gives the following sufficient condition for asymptotic stability of system (2).

Proposition 4.2 Let $P \in \mathbb{R}^{n \times n}$ be a positive-definite matrix and $\phi : \mathbb{R}^n \to \mathbb{R}$ a positive-definite function. Suppose $V(x) := x'Px$ is such that for any trajectory $x(\cdot)$ of (2)

$$\left(\frac{\partial V}{\partial x} f_{i(t)} \right) (x(t)) \leq -\phi(x(t)) \quad \text{for } t \geq 0.$$

Then the origin is asymptotically stable.

4.2 Problem Formulation

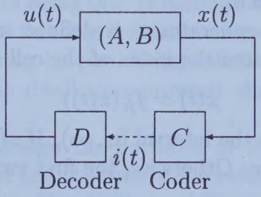


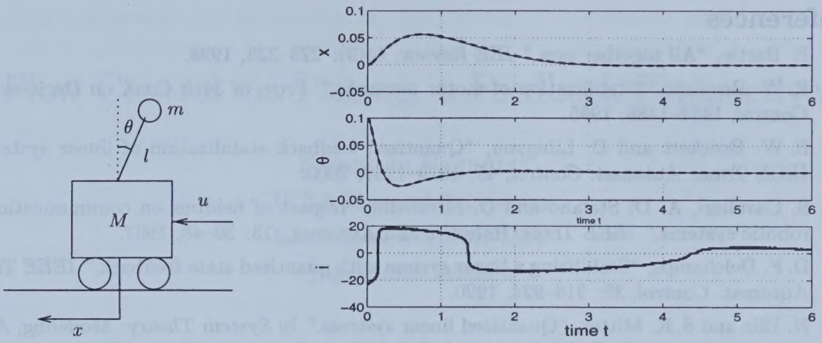
Figure 6: Dwell-time control scheme

The problem in this section is to design a dwell-time switched system that stabilizes a given linear time-invariant plant. Consider the feedback system in Fig. 6. The pair (A, B) represents the plant

$$\dot{x}(t) = Ax(t) + Bu(t),$$

where $x(t) \in \mathbb{R}^n$ is the state and $u(t) \in \mathbb{R}$ is the control input (single input is assumed). To avoid triviality, it is assumed that A is an unstable matrix, i.e., not all eigenvalues have negative real part, but that (A, B) is stabilizable.

The controller is termed a *dwell-time controller with dwell time* $T > 0$. Besides T, the controller is defined in terms of a logarithmic partition $\{\mathcal{X}_j\}_{j \in \mathbb{Z}_+}$ of $\mathbb{R}^n$ and a



(a) A cart with an inverted pendulum

(b) For $T = 0.012$

Figure 7:

corresponding set $\mathcal{U} = \{u_j\}_{j\in\mathbb{Z}_+}$ of control values. The decoder D in Fig. 6 is governed by the law

$$u(t) = u_{i(t)},$$

so the plant equation becomes

$$\dot{x}(t) = Ax(t) + Bu_{i(t)}.$$

This has the form (2), and therefore the switching logic of the preceding section defines the coder C mapping x to i.

We can now state the problem: Find a dwell time $T > 0$, a logarithmic partition $\{\mathcal{X}_j\}_{j\in\mathbb{Z}_+}$ of $\mathbb{R}^n$, and corresponding control values $\mathcal{U} = \{u_j\}_{j\in\mathbb{Z}_+}$ such that the closed-loop dwell-time switched system in Fig. 6 is asymptotically stable.

A solution to this problem is given in [10].

4.3 Numerical example

In this section, we apply the dwell-time switching controller to the familiar cart-pendulum system in Fig. 7(a). The cart is of mass $M = 2$ and position x, and force u is applied at one end by an actuator. On top of the cart is a pendulum of mass $m = 0.4$ kg, length $l = 0.5$ m, and angle θ rad.

Our result allows the dwell time to be less than 0.0138. We chose here $T = 0.012$, which resulted in a moderately large $\delta = 1.58$. The time response plots of x, θ, and i are given in solid lines in Fig. 7 for the initial condition $x(0) = \dot{x}(0) = \dot{\theta}(0) = 0$ and $\theta(0) = 0.1$ and $\alpha_0 = 1 \times 10^{-5}$. The ideal responses for full state feedback ($u = Kx$, no switching, no state partition), on which our design is based, are in dashed lines in the plots for x and θ; these overlap the states of the dwell-time controlled system. Signal i, transmitted over the channel, indicates the cell indices. It varies over about 40 levels. Note that in the simulation the infinite partition around the origin is impossible, so instead we used $\mathcal{X}_{00}^* = \cup_{j\le 0}(\mathcal{X}_j^+ \cup \mathcal{X}_j^-)$, corresponding to index 0 and control input 0. We can say from these plots that the proposed controller is efficient in partitioning the state space without causing much distortion in performance.

References

[1] E. Bartle, "All together now," *IEE Review*, 44(9): 223–225, 1998.

[2] R. W. Brockett, "Stabilization of motor networks," *Proc. of 34th Conf. on Decision and Control*, 1484–1488, 1995.

[3] R. W. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Trans. Automat. Control*, 45: 1279–1289, 2000.

[4] S. Cavalieri, A. Di Stefano and O. Mirabella, "Impact of fieldbus on communication in robotic systems," *IEEE Trans. Robotics and Automat.*, 13: 30–48, 1997.

[5] D. F. Delchamps, "Stabilizing a linear system with quantized state feedback," *IEEE Trans. Automat. Control*, 35: 916–924, 1990.

[6] N. Elia and S. K. Mitter, "Quantized linear systems," In *System Theory: Modeling, Analysis, and Control* (T. E. Djaferis and I. C. Schick eds.), Kluwer, Boston, 2000.

[7] N. Elia, "Design of hybrid systems with guaranteed performance," preprint, 2000.

[8] D. Hristu and K. Morgansen, "Limited communication control," *Systems Control Let.*, 37: 193–205, 1999.

[9] H. Ishii and B. A. Francis, "Stabilization with control networks," *Proc. UKACC Control 2000*, Cambridge, UK.

[10] H. Ishii and B. A. Francis, "Stabilizing a linear system by switching control with dwell time," submitted for publication, 2000.

[11] J. A. Janét, W. J. Wiseman, R. D. Michelli, A. L. Walker, M. D. Wysochanski and R. Hamlin, "Applications of control networks in distributed robotic systems," *Proc. Conf. on Syst., Men and Cybernetics*, 3365–3370, 1998.

[12] R. Jurgen, "Coming from Detroit: networks on wheels," *IEEE Spectrum*, 23(6): 53–59, 1986.

[13] H. K. Khalil, *Nonlinear Systems, 2nd Ed.*, Prentice-Hall, Uppersaddle River, N.J., 1996.

[14] P. P. Khargonekar and A. B. Özgüler, "Decentralized control and periodic feedback," *IEEE Trans. Automat. Control*, 39: 877–882, 1994.

[15] W. Lawrenz, *CAN System Engineering – From Theory to Practical Applications*, Springer, New York, 1997.

[16] A. S. Morse, "Control using logic-based switching," In *Trends in Control* (A. Isidori ed.), 69–114, Springer, London, 1995.

[17] G. N. Nair and R. J. Evans, "Stabilization with data-rate-limited feedback: tightest attainable bounds," *Systems Control Let.*, 41: 49–56, 2000.

[18] P. Pleinevaux and J.-D. Decotignie, "Time critical communication networks: fieldbuses," *IEEE Network*, 2(3): 55-63, 1988.

[19] R. S. Raji, "Smart networks for control," *IEEE Spectrum*, 31(6): 49–55, 1994.

[20] A. van der Schaft and H. Schumacher, *An Introduction to Hybrid Dynamical Systems*, Springer, London, 2000.

[21] D. D. Siljak, *Decentralized Control of Complex Systems*, Academic Press, Boston, 1991.

[22] V. I. Utkin, *Sliding Modes in Control and Optimization*, Springer, Berlin, 1992.

[23] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints I: state estimation problems," *IEEE Trans. on Autom. Control*, 42: 1294–1299, 1997.

[24] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints II: stabilization with limited information feedback," *IEEE Trans. on Autom. Control*, 44: 1049–1053, 1999.

Flow Control via Pricing: a Feedback Perspective

Fernando Paganini*
UCLA Electrical Engineering
Los Angeles, CA 90095-1594.
paganini@ee.ucla.edu

Abstract

This paper discusses network flow control from the perspective of multivariable feedback theory. We build on recent research on flow control based on pricing signals, which has cast these problems in terms amenable to mathematical analysis, and more specifically we focus on the control algorithm of [8]. After reviewing our recent result regarding the stability of this method, we move on to issues of performance and delay robustness. These are studied with tools of linear loop-shaping, which exhibit the main tradeoffs involved in parameter selection for these algorithms. An illustrative example is provided.

1 Introduction

Feedback mechanisms are of widespread use for controlling flows in networks such as the Internet; indeed some form of feedback is inevitable in order for resources to be efficiently used in the presence of high variability and uncertainty. However, feedback theory is rarely invoked for such designs: at first sight, practical congestion control mechanisms as those in TCP and its variants [3, 2] appear far removed from mathematically grounded control.

Recent research (see [1, 4, 5, 7, 8] and references therein) in flow control based on propagation of price signals has, however, strikingly closed this gap and opened the door for a mathematical theory of network flow control. These models can be used both to interpret the mechanism implicit in existing protocols [4, 9], and to propose alternatives based on more explicit price signaling. In particular, Kelly *et al.* [4, 5] have developed a framework with suggestive economic theory interpretations that leads to two dual families of flow control algorithms. Employing continuous time models, analytical results are obtained regarding equilibrium fairness, stability, and convergence rates for these systems, as well as studies of stochastic disturbances and time delays.

A related approach has been developed by Low and co-workers [7, 8], based on discrete-time models. In it, prices are interpreted as Lagrange multipliers for the optimization of total source *utility*, subject to link capacity constraints. This interpretation leads to convergence proofs for a first-order version of the price update law [7]. A second-order version was later proposed in [8] to address some drawbacks of the first order algorithm. This alternative has satisfactory performance in simulations, but analytical results have been more difficult to obtain. Very recently [11], employing a continuous-time model and a Lyapunov argument, we have been able to prove stability for this case. An outline of this result is provided in Section 3 of this paper.

---

*Research supported by NSF CAREER Award ECS-9875056, and the David and Lucille Packard Foundation. The author thanks Steven Low for his guidance in this topic.

The main thrust of the present paper is, however, to move beyond stability and apply to this problem the perspective of the multivariable feedback designer. Here, as in other applications, the main concern is the fundamental tradeoff between dynamic *performance* (transient response, tracking capability, etc.) and *robustness* to uncertain or under-modeled effects (e.g. time delays). In this regard, we show in Section 4, working with linearized models, that a loopshaping point of view can provide insight on the design tradeoffs involved in parameter choices for the algorithm of [8]. Section 5 contains a simulation example that illustrates these findings in the nonlinear setting.

2 Problem Formulation

We are concerned with a system of L communication links shared by a set of K sources. The routing matrix R, of dimensions $L \times K$, is defined by

$$R_{lk} = \begin{cases} 1 & \text{if source } k \text{ uses link } l \\ 0 & \text{otherwise} \end{cases}$$

- **For each link l we have:**
    - the capacity c_l;
    - the aggregate rate y_l of all flows through the link;
    - the backlog b_l;
    - the price signal p_l.
- **For each source k we have:**
    - The source rate x_k;
    - The aggregate price q_k of all links used by source k.

We use vector notation to collect the above variables across all links or sources; thus we define $c, y, b, p \in \mathbb{R}^L$, and $x, q \in \mathbb{R}^K$. The following relationships are immediate[1]:

$$y = Rx, \qquad q = R^T p. \tag{1}$$

It is assumed that links are able to measure their own aggregate rate, and that sources are fed back their aggregate price; see [5, 8] for discussions on implementation of such mechanisms. For the moment we also assume that propagation of rates and prices is instantaneous; we later discuss the effect of delay.

With the information given so far, we have depicted in Figure 1 a generic feedback structure for flow control based on price propagation; what remains to be specified is:

- How the links fix their prices based on link utilization.
- How the sources fix their rates based on their aggregate price.

---

[1] R^T is the matrix transpose of R.

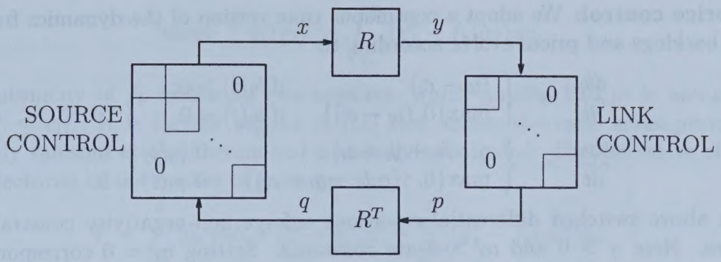


Figure 1: General flow control structure based on pricing signals.

These operations are up to the designer, but have a main restriction: both must be *decentralized*, as indicated in the figure by a block diagonal structure. For instance the source rate x_k can only depend on the corresponding aggregate price q_k. We are thus in the realm of decentralized feedback control. The objective of this feedback is to allow for flows to dynamically adapt to changing conditions in traffic demand, link capacity, routing, etc. Key design considerations are thus to ensure dynamic stability and regulation of these systems around equilibria that satisfy some desirable static properties (full resource utilization, fairness among sources).

In this context, Kelly *et al.* [5] have proposed two flow control methods:

(i) The "primal" algorithm uses a first-order continuous-time model for the source rate control, and a static law for the price update.

(ii) The "dual" method uses a first-order dynamics for the price update, and a static law for source rate selection.

We will not describe the specific laws here, but indicate that [5] describes fairness properties of the resulting equilibria, and proves that these are global attractors by means of a Lyapunov argument. These equilibria can also be interpreted in economic terms as balancing network supply with source demand, expressed in terms of a *utility* function.

In this paper we will focus on the approach of Low and coworkers [7, 8], which is similar to the dual method above, in particular it is based on a static law for source control, and a dynamic price control law at the links.

Source rate control. According to [7, 8], for a given total price q_k, the sources pick the rate that maximizes

$$\mathcal{U}_k(x_k) - x_k q_k$$

over x_k, where $\mathcal{U}_k(x_k)$ is the source utility function, assumed to be strictly concave[2]. Assuming $\mathcal{U}_k(\cdot)$ is differentiable, the maximum is achieved at $x_k = \mathcal{U}_k'^{-1}(q_k)$, where $\mathcal{U}_k'^{-1}$ is the inverse function of the derivative of $\mathcal{U}_k$. We denote henceforth

$$f_k(q_k) := \mathcal{U}_k'^{-1}(q_k).$$

Notice that $\mathcal{U}_k'$ is strictly decreasing in $x_k > 0$, hence f_k is a strictly monotone decreasing function of q_k. In vector notation, we summarize the above equations for source rates as

$$x = f(q). \tag{2}$$

---

[2] [7] allows for the inclusion of maximum and minimum constraints for x_k, for simplicity we will not impose those here (as, for instance, in logarithmic utility functions $\mathcal{U}_k(x_k) = w_k \log(x_k)$).

Link price control: We adopt a continuous time version of the dynamics from [8]; for each l, backlogs and prices evolve according to

$$\frac{db_l}{dt} = \begin{cases} (y_l - c_l) & \text{if } b_l(t) > 0; \\ \max\{0, (y_l - c_l)\} & \text{if } b_l(t) = 0. \end{cases} \tag{3}$$

$$\frac{dp_l}{dt} = \begin{cases} \gamma(\alpha_l b_l + y_l - c_l) & \text{if } p_l(t) > 0; \\ \max\{0, \gamma(\alpha_l b_l + y_l - c_l)\} & \text{if } p_l(t) = 0. \end{cases} \tag{4}$$

The above switched differential equations enforce non-negativity constraints in all variables. Here $\gamma > 0$ and $\alpha_l > 0$ are constants. Setting $\alpha_l = 0$ corresponds to the first-order price update law in [7]; in this case prices are proportional to backlogs, which may lead to an undesirable equilibrium backlog, hence the use of $\alpha_l > 0$ in [8].

Let (b^*, p^*) be an equilibrium of the above system. We also use the notation $q^* = R^T p^*$ for the equilibrium source prices, $x^* = f(q^*)$ for the equilibrium source rates, and $y^* = Rx^*$ for the equilibrium link rates. It is not difficult to see that we must have $b^* = 0$. Indeed, if $b_l^* > 0$ then we would have $y_l^* = c_l$ so $\dot{p}_l > 0$, which contradicts equilibrium. Now p^* need not be zero, indeed its nonzero components correspond to links where $y_l \equiv c_l$, i.e. where the capacity constraint is active (bottleneck links).

The relation of this algorithm with optimization is explained in detail in [7], but we briefly outline it here. The key observation is that an equilibrium point p^*, x^* satisfying the equations (1-4) will be a saddle point of the optimization

$$\min_{p \geq 0} \max_{x} \left(\sum_{k=1}^{K} \mathcal{U}_k(x_k) + p^T(c - Rx) \right).$$

This is the Lagrangian dual of the convex program of maximizing the overall utility $\sum_{k=1}^{K} \mathcal{U}_k(x_k)$, subject to link capacity constraints $Rx \leq c$.. Furthermore, the first-order link price algorithm of [7], corresponding to $\alpha_l = 0$, can be interpreted as a gradient algorithm for the above maximization in p; this is used to establish convergence. No such interpretation is available in the second-order case of [8] we are considering.

It follows from duality theory that x^* must be the unique global optimum of the latter problem; therefore y^*, q^* are also unique. p^* need not be unique, because in general the capacity constraints might not be independent. To simplify the further development and obtain a unique equilibrium price, we assume from now on that the matrix R is of full row rank. Thus, given a vector q of aggregate source prices, there is a unique p satisfying $q = R^T p$. Some comments on generalizing this assumption are given in [11].

3 Stability

In this section we outline a proof of global stability for the price update law of (3-4); this supplements available results for the systems in [5, 7]. We state the following result:

Theorem 1. *Given the system (1-4), assume $f_k(q_k)$ is strictly decreasing in $q_k > 0$, and that R is of full row rank. Then the unique equilibrium point $b^* = 0$, p^* is globally asymptotically stable.*

A detailed proof is given in [11]. Here we will briefly outline the argument, which is based on the Lyapunov function

$$V(b, p) = \sum_{l=1}^{L} [\alpha_l \gamma \frac{b_l^2}{2} + (c_l - y_l^*)p_l] + \sum_{k=1}^{K} \phi_k(q_k),$$

$$\text{where} \quad \phi_k(q_k) := \int_{q_k^*}^{q_k} (x_k^* - f_k(\sigma))d\sigma.$$

The monotonicity of f_k makes ϕ_k non-negative, which implies $V(b,p)$ is non-negative (notice $y_l^* \leq c_l$). It is further argued in [11] that under the rank assumption on R, $V(b,p)$ only vanishes at equilibrium and is radially unbounded. The derivative of $V(b,p)$ along trajectories of the system is given by

$$\dot{V} = \sum_{l=1}^{L}[\alpha_l \gamma b_l \dot{b}_l + (c_l - y_l^*)\dot{p}_l] + \sum_{k=1}^{K}(x_k^* - f_k(q_k))\dot{q}_k.$$

Now observe that last term above can be written as

$$\sum_{k=1}^{K}(x_k^* - x_k)\dot{q}_k = (x^* - x)^T \dot{q} = (x^* - x)^T R^T \dot{p} = (y^* - y)^T \dot{p} = \sum_{l=1}^{L}(y_l^* - y_l)\dot{p}_l, \tag{5}$$

and thus

$$\dot{V} = \sum_{l=1}^{L}[\alpha_l \gamma b_l \dot{b}_l + (c_l - y_l^*)\dot{p}_l + (y_l^* - y_l)\dot{p}_l] = \sum_{l=1}^{L} \nu_l,$$

where we have denoted $\nu_l := \alpha_l \gamma b_l \dot{b}_l + (c_l - y_l)\dot{p}_l$.

Consider the case where $b_l > 0$, $p_l > 0$; then the dynamics (3-4) give

$$\nu_l = \alpha_l \gamma b_l (y_l - c_l) + (c_l - y_l)\gamma(\alpha_l b_l + y_l - c_l) = -\gamma(y_l - c_l)^2 \leq 0.$$

The limiting cases $b_l = 0$, $p_l = 0$ are more delicate, but a detailed study in [11] shows that indeed $\nu_l \leq 0$ at all times. Therefore our Lyapunov function is decreasing along trajectories of the system. Finalizing the proof of global asymptotic stability involves ensuring that $\dot{V} \equiv 0$ can only occur at equilibrium, and invoking Lasalle's invariance principle (see [6]) to show trajectories must converge to equilibrium over time. For full details on this last step see [11].

This theorem implies, under mild assumptions, that the flow control method proposed in [8] is sound and will, in the continuum limit, converge asymptotically to a desirable equilibrium where the overall utility is maximized, and backlogs are cleared.

4 Performance and Robustness

Global stability is a fundamental property, and it is fortunate to be able to establish it with such generality; however it does not suffice to guarantee a feedback system behaves in a satisfactory way. As in other feedback design problems, key questions are:

(i) What is the dynamic performance of this system? There are many ways to specify this, but mainly one wants to know that in addition to "eventually" settling down to equilibrium, the system is capable of quickly reacting to changing conditions. Of relevance here are the transient response from non-equilibrium initial conditions (these arise due to changes in network users, links, or routing), and the ability to track variations in link capacity c (e.g. in an Available Bit Rate scenario).

(ii) What is the robustness of this system to deviations from the idealized model? There are again many such deviations, many of which can be modeled as stochastic disturbances around equilibrium. From a stability point of view, however, the key concern is the effect of system *delay*, which arises both in propagating prices and rates and also is inherent in discrete-time price update implementations.

Some of these issues have been studied in [5] for the pricing algorithms considered there. Here we will explore some of these questions for the algorithm in [8], and cast them in the language of multivariable feedback control theory. We will work with a *linearized* model of the dynamics around the equilibrium. Let $\tilde{x}$, $\tilde{y}$, etc, denote the deviations of the system variables from their equilibrium values x^*, y^*, and so on. We also allow fluctuations $\tilde{c}$ in the link capacities. Then:

- Linearizing the source rate function $x = f(q)$ around the point (x^*, q^*), we write $\tilde{x} \approx -D\tilde{q}$, where
$$D = -\mathrm{diag}(f_k'(q_k^*)) = -\mathrm{diag}\left(\frac{1}{\mathcal{U}_k''(x_k^*)}\right).$$
Here and in the sequel, diag$(\cdots)$ denotes a diagonal matrix with the enclosed diagonal elements. We have used above the fact that f_k is $(\mathcal{U}_k')^{-1}$. The sign convention is for convenience, making D a positive definite matrix (since $\mathcal{U}_k$ is concave).
- We consider the linear approximation
$$\frac{d\tilde{b}_l}{dt} = (\tilde{y}_l - \tilde{c}_l) \tag{6}$$
$$\frac{d\tilde{p}_l}{dt} = \gamma(\alpha_l \tilde{b}_l + \tilde{y}_l - \tilde{c}_l) \tag{7}$$
to the price/backlog dynamics. Clearly, (7) is a valid local model for bottleneck links, where the equilibrium $p_l^* > 0$. The argument is less clear for (6), since the equilibrium of (3) occurs at the point of discontinuity; one case in which it would apply is if the b_l are taken to measure the deviation from a nonzero backlog b_{l0}, as is also considered in [8]. Given these limitations, as well as the inherent locality of any linear model, our analysis to follow does not carry the same weight as the earlier stability result. Nevertheless we will see below that this first-cut linear model can have significant predictive power.

With this model, we can represent the linearized dynamics by the block diagram on the top of Figure 2; in it, s is the usual Laplace variable, and for simplicity we have assumed that the α_l have the same value α across all links.

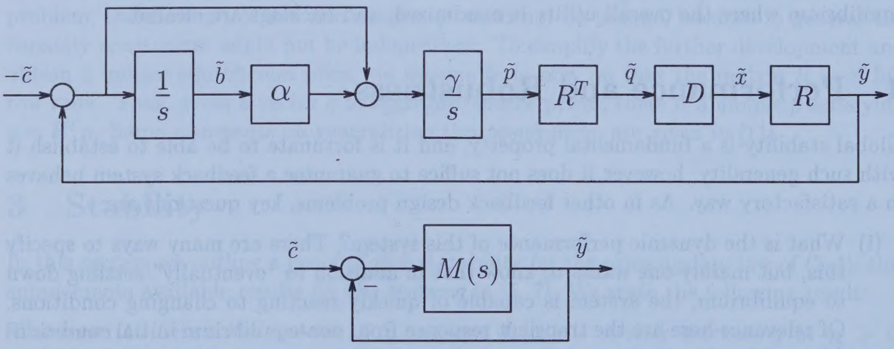


Figure 2: Linearized model and standard unity feedback form

The bottom figure is obtained from the above by absorbing the negative sign and defining the loop gain transfer function matrix
$$M(s) = \frac{\gamma}{s}\left(1 + \frac{\alpha}{s}\right) RDR^T.$$

In this way we are led to a standard unity feedback configuration; we can now bring in standard tools from multivariable control, to analyze:

1. The transient response. In particular, the closed loop modes are given by the roots of $\det(I + M(s)) = 0$.

2. The tracking performance: i.e. how well the link utilization tracks link capacity variations. Here we can look at the singular values of the sensitivity function
$$S(j\omega) = (I + M(j\omega))^{-1}$$
which maps $\tilde{c}$ to tracking error, and study the range of ω for which $S(j\omega)$ is small.

3. Stability margin to delay. One can insert delays of the form $e^{-\tau_i s}$ to loop channels and study bounds on delays consistent with stability.

These tasks are greatly simplified by the observation that the loop transfer function is easily diagonalized: indeed the matrix RDR^T is symmetric, positive definite and hence can be written as $U^T \Lambda U$, where U is a unitary matrix and Λ is the diagonal matrix of positive eigenvalues $\lambda_1, \dots \lambda_L$. Given this fact, we have
$$M(s) = U^T \mathrm{diag}\Big(m_1(s), \dots, m_L(s)\Big) U,$$
with $m_i(s) = \frac{\gamma\lambda_i}{s}\left(1 + \frac{\alpha}{s}\right)$. This means that the multivariable feedback can be easily studied by L single loops of loop-gain $m_i(s)$, leading to the following conclusions:

1. The closed loop modes are the roots of $s^2 + (\gamma\lambda_i)s + \gamma\lambda_i\alpha = 0$, for $i = 1 \dots L$. It is easily verified that the dominant pole of the above equation can have negative real part no faster than $\gamma\lambda_i/2$; therefore the decay exponent of the overall transient dynamics is limited by $\gamma\lambda_{\min}/2$, where $\lambda_{\min}$ is the smallest eigenvalue of RDR^T.

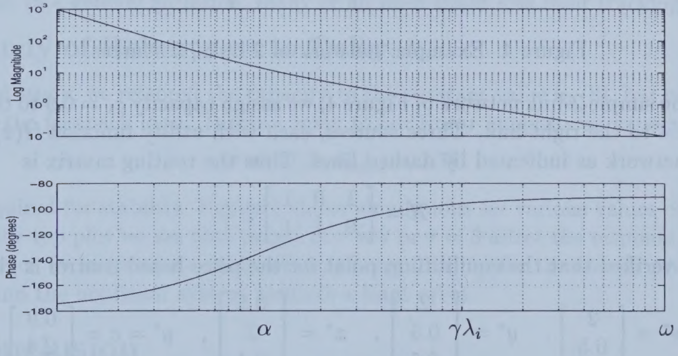


Figure 3: Bode plot of $m_i(j\omega)$

2. The sensitivity function diagonalizes as $S(j\omega) = U\mathrm{diag}\Big((1 + m_i(j\omega))^{-1}\Big)U^T$, so its singular values are determined by the "loopshape" functions $m_i(j\omega)$. In Figure 3 we sketch its Bode plot for the case $\alpha << \gamma\lambda_i$. The frequency range for which $|m_i(j\omega)| >> 1$ is associated with good tracking performance. Since the cross-over frequency is about $\gamma\lambda_i$ in this case, once again we find the figure of merit $\gamma\lambda_{\min}$ as a measure of the bandwidth over which our system exhibits good tracking.

3. Finally, we turn to the issue of robustness to delay. The analysis is simplest for a common delay $e^{-\tau s}$ in all loops; this directly translates to a modified loopshape function $m_i(j\omega)e^{-\tau j\omega}$; for this to yield stability, in classical terms we must have a *phase margin* greater than $\omega_c \tau$ in the original Bode plot, where ω_c is the crossover frequency.

In fact the best phase margin is about 90 degrees, achieved when $\alpha << \gamma\lambda_i$, which motivates our earlier choice. Then $\omega_c \approx \gamma\lambda_i$ and the stability condition is approximately

$$\tau < \pi/(2\gamma\lambda_i).$$

Here the condition is more stringent for the *largest* eigenvalue $\lambda_{\max}$ of RDR^T, and is *inversely* proportional to γ.

We summarize the conclusions of our linear analysis:

- The parameter γ directly affects dynamic performance, but inversely affects delay robustness; a judicious design must balance this tradeoff.
- A similar role is played by the matrix RDR^T, which represents the sensitivity of link rates to price increases. Note, however, that while its *minimum* eigenvalue dictates performance, its *maximum* eigenvalue impacts delay stability. Hence, for an ill-conditioned matrix RDR^T the design tradeoff will become more difficult; this observation is well known in multivariable feedback design problems (see, e.g. [10]).
- For the parameter α, robustness to delay indicates the use of $\alpha << \gamma\lambda_i$.

5 Example

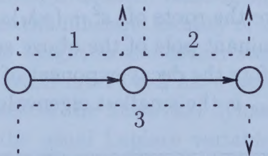


Figure 4: Example: network of 2 links, 3 sources

Consider the simple 2-link network of Figure 4; we assign capacity $c_1 = 0.9$ to the left link, and $c_2 = 2.4$ to the right link. Three sources, each with utility function $\mathcal{U}(x) = \log(x)$, share the network as indicated by dashed lines. Thus the routing matrix is

$$R = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

It is easily verified that the equilibrium point for the price-based control is given by

$$p^* = \begin{bmatrix} 2 \\ 0.5 \end{bmatrix}, \quad q^* = \begin{bmatrix} 2 \\ 0.5 \\ 2.5 \end{bmatrix}, \quad x^* = \begin{bmatrix} 0.5 \\ 2 \\ 0.4 \end{bmatrix}, \quad y^* = c = \begin{bmatrix} 0.9 \\ 2.4 \end{bmatrix}. \tag{8}$$

The linearized matrix

$$D = -\mathrm{diag}\left(\frac{1}{\mathcal{U}''(x_1^*)}, \frac{1}{\mathcal{U}''(x_2^*)}, \frac{1}{\mathcal{U}''(x_3^*)}\right) = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 0.16 \end{bmatrix}$$

leads to $RDR^T = \begin{bmatrix} 0.41 & 0.16 \\ 0.16 & 4.16 \end{bmatrix}$, with eigenvalues $\lambda_{\min} = 0.4$, $\lambda_{\max} = 4.16$.

We see that we are in an ill-conditioned case, with $\lambda_{\max}/\lambda_{\min} \approx 10$.

5.1 Study of Tracking Performance

We consider the price dynamics of (3-4) with parameters $\gamma = 0.1$, $\alpha = 0.01$. Note that $\gamma\lambda_{\min} = 0.04$, so we are in the situation of Figure 3, and in particular, the cross-over frequency of $m_1(j\omega)$ is $\omega_{c1} \approx 0.04$. Our linear analysis predicts that variations in capacity will be tracked well up to this bandwidth, but not beyond.

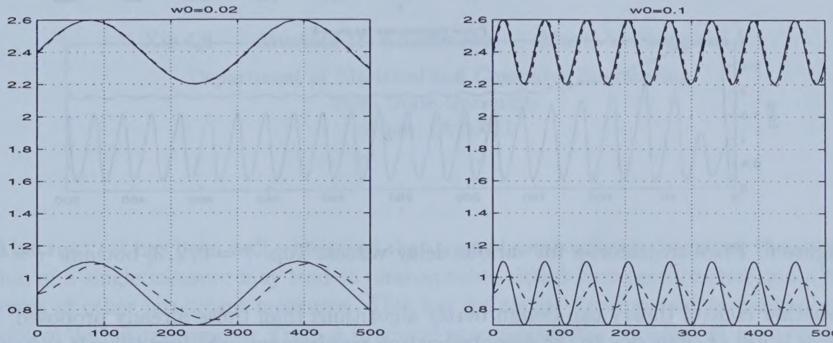


Figure 5: Link capacities (solid) and flows (dashed). Left: $\omega_0 = 0.02$; right: $\omega_0 = 0.1$

We simulated the nonlinear dynamics (1-4) using the Matlab stiff solver 'ode15s', and introducing the sinusoidal variations $\tilde{c}_l(t) = 0.2\sin(\omega_0 t)$, to the capacity values from (8), for several values of ω_0. Figure 5 presents the results for $\omega_0 = 0.02$ and $\omega_0 = 0.1$. We confirm the prediction that at least one of the link flows loses the ability to track the network capacity as frequency grows beyond the critical value of $\omega_0 = 0.04$. The fact that the other flow is able to track closely corresponds here to the matrix RDR^T being close to diagonal; in a general situation, many or all links could lose their tracking capability.

5.2 Study of Robustness to Delay

In a second study, we restored the capacity to the constant value in (8), but now inserted a delay in the feedback loop. Turning again to our linear analysis, we expect that

$$\tau < \pi/(2\gamma\lambda_{\max}) = 3.77$$

will be required for stability. Figure 6 shows simulations for various values of the delay.

From the top plot we see that delays of $\tau = 2$ or $\tau = 3$ affect the response, but retain stability; once we move to $\tau = 4$, however, the equilibrium for the second price becomes unstable and the nonlinear system exhibits a limit cycle.

6 Conclusion

We have explored the application of tools from feedback theory to flow control algorithms based on pricing. Our investigation with simple linear models reveals basic design tradeoffs involved when selecting control parameters, e.g. the price update parameter γ in [8]. An interesting finding is that an ill-conditioned matrix RDR^T can severely affect these dynamic tradeoffs. This factor should thus be considered, in addition to other considerations such as fairness, in decisions on routing and utility function selection.

We have focused on the control law from [8], but in principle similar analysis is possible for other methods with the general structure of Figure 1. A natural question

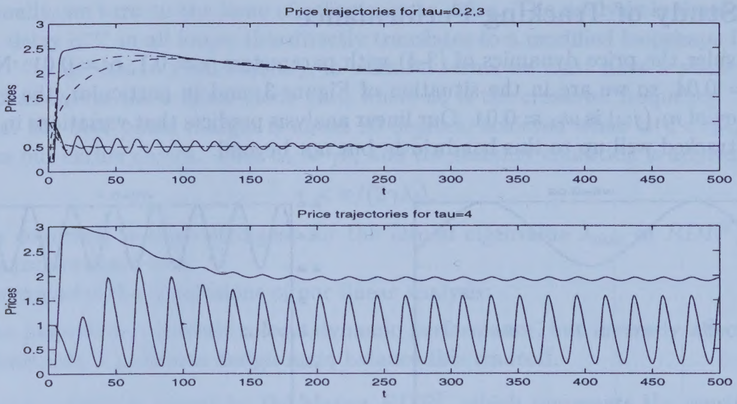


Figure 6: Price trajectories for various delay values. Top: $\tau = 0, 2, 3$; bottom: $\tau = 4$.

is whether control theory can design better algorithms than those already proposed; the answer is not obvious due to the decentralization requirement, which is difficult to handle in control synthesis methods; pursuing it will be a topic of future research.

References

[1] C. Courcoubetis, V. Siris, and G.D.Stamoulis, "Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks", *Proc. Globecom '96.*

[2] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", *IEEE/ACM Trans. on Networking,* 1(4):397-413, Aug. 1993.

[3] V. Jacobson, "Congestion avoidance and control", *Proc. ACM SIGCOMM '88.*

[4] F. P. Kelly, "Mathematical modeling of the Internet", Fourth International Congress on Industrial and Applied Mathematics, Edinburgh, Scotland, July 1999.

[5] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability". *Jour. Oper. Res. Soc.*, 49(3), pp. 237-252, 1998.

[6] H.K. Khalil. *Nonlinear Systems.* Prentice-Hall, 1996.

[7] S. H. Low and D. E. Lapsley, "Optimization flow control – I: basic algorithm and convergence" *IEEE/ACM Trans. on Networking,* Vol 7, No 6. Dec 1999.

[8] S. Athuraliya and S. H. Low, "Optimization flow control – II: random exponential marking" *preprint,* http://www.ee.mu.oz.au/staf/slow/research, May 2000.

[9] S. Low, L. Peterson and L. Wang, "Understanding TCP Vegas: Theory and Practice", *preprint,* http://www.ee.mu.oz.au/staf/slow/research, Feb. 2000.

[10] M. Morari and E. Zafiriou. *Robust Process Control.* Prentice Hall, 1989.

[11] F. Paganini, "On the stability of optimization-based flow control", submitted to the 2001 American Control Conference. http://www.ee.ucla.edu/~paganini/

A New Approach to Multi-Objective Controller Synthesis[*]

Xin Qi Mustafa H. Khammash Murti V. Salapaka [†]

Department of Electrical and Computer Engineering
Iowa State University
Ames, IA 50011

Abstract It has been well recognized that control system design based on the optimization of a single measure may lead to unacceptable closed-loop performance quantified in terms of other important measures. This has led to the development of multi-objective control design techniques. From a practical standpoint, the problem of multi-objective design usually involves a control design problem with mixed frequency-domain and time-domain specifications. Many of these requirements are adequately captured using $\mathcal{H}_2$, $\mathcal{H}_\infty$, and ℓ_1 norms. One important time-domain constraint that cannot be naturally captured using these system norms is the transient response to reference input signals or disturbances (e.g. steps). Most current multi-objective design approaches address only a subset of these constraints. We provide an overview of such methods and then present a control design methodology that incorporates all the above specifications in one unified multi-objective framework. We then provide a solution for the resulting multi-objective problem.

Keywords $\ell_1/\mathcal{H}_2/\mathcal{H}_\infty$, time-domain constraint (TDC), multi-objective control, robust optimal control, semidefinite programming

1 Introduction

The last ten years has witnessed a rapid growth of research effort addressing multi-objective control problems. The fundamental reason behind this is that the performance requirements that must be satisfied by advanced engineering system are becoming more stringent, and involve multiple, often conflicting, objectives. One example where these multi-objective concerns exist naturally is the suspension control system design problem described in [1]. In this system, suspensions are designed to achieve several conflicting goals that can be naturally expressed in terms of $\mathcal{H}_2$, ℓ_1, and $\mathcal{H}_\infty$ norm-based objectives. The suspension controller synthesis can then be transformed to the problem of searching a suitable tradeoff among the above three norm-based objectives. This example, as well as many others suggest the need for a design methodology that can simultaneously

---

[*] This research was supported by the National Science Foundation under Grants No. ECS-9457485 and ECS-9733802.

[†] Author's Emails: `xinqi@iastate.edu, khammash@iastate.edu, murti@iastate.edu`.

handle multi-objectives typically involving the time and frequency domains. Mathematically, these constraints result in an optimization problem with various norm constraints including the ℓ_1, $\mathcal{H}_2$, and $\mathcal{H}_\infty$ norms, as well as other constraints that arise from transient response specification in the case when reference input signals and/or disturbances are present. Most existing multi-objective approaches address only a subset of the above performance objectives, and focus mainly on objectives that can be captured by closed-loop system norms. Examples include $\mathcal{H}_2/\mathcal{H}_\infty$, $\mathcal{H}_2/\ell_1$, $\ell_1/\mathcal{H}_\infty$, etc. The nature of the proposed solution techniques to many of these mixed-objective problems makes it difficult to generalize these techniques to include additional objectives. The difficulty stems from the different nature of the system norms under consideration and the difference in the tools that have been traditionally developed to minimize each of these norms. In this paper we show that the Scaled-Q method, initially proposed in [2] and later adapted to solve the $\mathcal{H}_2/\ell_1$ problem in and [3] and [4], can be generalized to handle multi-objective control problems having ℓ_1, $\mathcal{H}_2$, and $\mathcal{H}_\infty$ norm constraints as well as time-domain template constraints on the transient response to command reference inputs and/or disturbances.

The paper is organized as follows. In Section 2, we introduce the notation to be used. In Section 3, we formulate the setup of the multi-objective control problem to be used in the paper. In Section 4, we survey various approaches addressing the extensively studied multi-objective problems. In Section 5, we introduce a new effective methodology which solves the general multi-objective problem involving all the four types of objectives listed above. Finally, in Section 6, we present a brief conclusion.

2 Notation

ℓ_1	The Banach space of right sided absolutely summable real sequences with the norm given by $\|\|x\|\|_1 := \sum_{k=0}^{\infty} \|x(k)\|$.
$\ell_1^{m\times n}$	The Banach space of matrix-valued right-sided real sequences with the norm given by $\|\|x\|\|_1 := \max_{1\le i\le m} \sum_{j=1}^{n} \|\|x_{ij}\|\|_1$, where $x \in \ell_1^{m\times n}$ is the matrix (x_{ij}) and each x_{ij} is in ℓ_1.
ℓ_2	The Banach space of right-sided square summable real sequences with the norm given by $\|\|x\|\|_2 := \left(\sum_{k=0}^{\infty} x^2(k)\right)^{1/2}$.
$\hat{x}(\lambda)$	The λ transform of a right-sided real sequence $x = (x(k))_{k=0}^{\infty}$ defined as $\hat{x}(\lambda) := \sum_{k=0}^{\infty} x(k)\lambda^k$.
$\mathcal{H}_2$	The isometrically isomorphic image of ℓ_2 under the λ transform with the norm given by $\|\|\hat{x}(\lambda)\|\|_{\mathcal{H}_2} = \|\|x\|\|_2$.
$\mathcal{H}_\infty^{m\times n}$	The space of complex-valued matrix functions that are analytic in the open unit disc and bounded on the unit circle with the norm defined by $\|\|\hat{x}\|\|_{\mathcal{H}_\infty} := \operatorname{ess\,sup}_\theta \sigma_{max}[\hat{x}(e^{i\theta})]$.

3 Formulation of Problem

In this section, we formulate the multi-objective problem addressed in the remainder of this paper. Consider the system shown in Figure 1, where G is the generalized linear time-invariant plant, K is the controller. w, z, u, and v represent the exogenous input, regulated output, control input, and measured output, respectively. Let n_{w_i}, n_{z_i}, n_u and n_y denote the dimensions of w_i, z_i, n_u and n_y. Then $w := [w_1 \dots w_I]^T$ and $z := [z_1 \dots z_I]^T$ are of dimensions $n_z := \sum_{i=1}^{I} n_{w_i}$ and $n_w := \sum_{i=1}^{I} n_{z_i}$ respectively.

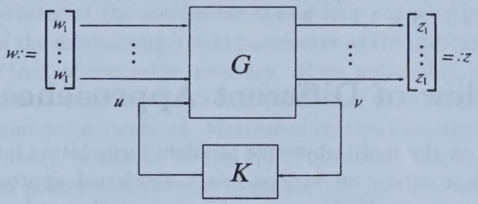


Figure 1: Closed-loop system

Let $\hat{R}$ denote the closed-loop transfer matrix from w to z and $\hat{R}^i$ $(i = 1, \dots, I)$ the closed-loop transfer matrix from w_i to z_i. Given the system described above, the various frequently encountered multi-objective problems (e.g. $\mathcal{H}_2/\mathcal{H}_\infty$, $\ell_1/\mathcal{H}_\infty$) can be formulated by imposing specific performance criteria over certain signal pairs (w_i, z_i), which can be ultimately reduced to frequency- and/or time-domain specifications imposed upon the admissible closed-loop transfer matrix $\hat{R}^i$. This will lead to a search over all the stabilizing controllers to ascertain an optimal or suboptimal solution to the problem of interest. The multi-objective problems frequently encountered in the literature usually take one of the two formalisms described below:

1. Find a stabilizing controller which minimizes a nonnegative linear combination of the closed-loop norms of interest:

$$\inf_{K\, stabilizing} c_1 \|\hat{R}^1(K)\|_{g_1} + c_2 \|\hat{R}^2(K)\|_{g_2} \tag{1}$$

   where $c_i \geq 0$, g_i is ℓ_1, $\mathcal{H}_2$, or $\mathcal{H}_\infty$. This type of problem will be referred as the combination problem in this paper. Note that all the solutions obtained by varying ratios between c_1 and c_2 are just the set of all Pareto optimal solutions of (1).

2. Find a stabilizing controller that minimizes $\|\hat{R}^1(K)\|_{g_1}$ while satisfying certain constraint on the closed-loop transfer matrix $\hat{R}^2(K)$, i.e. :

$$\inf_{K\, stabilizing} \{\|\hat{R}^1(K)\|_{g_1} : \ subject\ to\ \|\hat{R}^2(K)\|_{g_2} \leq \gamma\}. \tag{2}$$

   This type of problem will be referred as the mixed problem in the paper.

From the engineering point of view, the combination problem is naturally posed to study the performance tradeoff between the two different performance objectives. The mixed problem furnishes a framework to synthesize controller(s) achieving the best achievable performance in one channel while maintaining a specific constraint level on the nominal performance of another closed-loop channel. It is clear that the above two problems are closely related. At the same time, the solution of one problem is not directly helpful for obtaining the solution to the other. For this reason, we will address the two problems separately while keeping in mind the relation between them.

In this paper, we make use of the Youla parametrization (also referred as the Q-parametrization), where the set of all the achievable closed-loop maps is characterized as([10]):

$$\{\hat{R} = \hat{H} - \hat{U}\hat{Q}\hat{V} \text{ with } \hat{Q} \text{ stable}\}, \tag{3}$$

where $\hat{Q}$ is a free stable parameter, and the transfer matrices $\hat{H}$, $\hat{U}$, $\hat{V}$ are all stable.

Let H, U and V in the Youla parametrization (3) be partitioned into submatrices of compatible dimensions with the exogenous input component w_i and regulated output component z_i. Then the closed-loop transfer matrices from w_i to z_i can be expressed as $R^i(Q) = H^{ii} - U^i * Q * V^i, i = 1, \ldots, 7$.

4 An Overview of Different Approaches

In this section, based on the multi-objective problem formulation introduced in the last section, we shall make a survey on various newly-developed approaches frequently encountered in the literature (see [8] for more references on the work previous to 1997).

The multi-objective problems involving the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms have received considerable attention in the past decade. In particular, the past three years has seen a rapid growth of more advanced approaches to this type of problem. In [11]-[15], various approaches were proposed to compute and improve upper bounds to the $\mathcal{H}_2 + \mathcal{H}_\infty$ combination problem. In [16], an LMI-based approach was presented to compute a sequence of bounds that converges to the optimum monotonically from below. This complements the solution to the multi-objective $\mathcal{H}_2 + \mathcal{H}_\infty$ combination problem by furnishing a stopping criterion for the algorithms developed in [11]-[15]. In [17] and [18], an LMI-based approach and a randomized approach are proposed respectively to obtain upper bound(s) to the $\mathcal{H}_2/\mathcal{H}_\infty$ problem. So far it was not clear how to compute (nondecreasing) lower bound(s) to the $\mathcal{H}_2/\mathcal{H}_\infty$ mixed problem by using the LMI machinery.

In [10] and [19], linear programming and duality theory were used to solve this problem by approximating the $\mathcal{H}_\infty$ constraint with a finite set of linear constraints obtained by sampling the unit circle. It has been shown, however, that for a class of problems, this approximation may fail to converge even as the number of sampling points tend to infinity([7] and [20]).

In [21], the solution to the four-block $\ell_1/\mathcal{H}_\infty$ problem was obtained by solving a finite-dimensional convex optimization problem together with an unconstrained $\mathcal{H}_\infty$ problem. The essential idea in this approach is to replace $\mathcal{RH}_\infty$ by $\mathcal{RH}_{\infty,\delta}$ where the latter denotes the transfer matrix subspace of $\mathcal{RH}_\infty$ that have analytic continuation inside the disc of radius $\delta > 1$ (usually $\delta \approx 1$). Then the optimal performance μ of $\ell_1/\mathcal{H}_\infty$ mixed problem can be found by solving a sequence of modified problems whose optimum value μ_{δ_i} converges monotonically to μ from above. By constructing a procedure of computing lower bounds of μ, this approach can yield suboptimal solution with performance arbitrary close to the optimum. In [23], the existence of an optimal solution to the multi-block $\ell_1/\mathcal{H}_\infty$ problem is established. Moreover, in [23] it was shown that the optimal solution can be approximated arbitrarily closely by real-rational transfer matrices and more interesting from a computational viewpoint, a bound on the order of a ϵ-suboptimal solution is also given.

For multi-objective problems involving the $\mathcal{H}_2$ and ℓ_1 norms, two main lines of approach have been developed to obtain the solution. In [19] and [26], solutions to the $\mathcal{H}_2/\ell_1$ mixed problem were developed by using linear programming (LP) techniques combined with duality theory. Nevertheless since the set of achievable closed-loop maps is characterized by using zero interpolation, this line of approaches would usually lead to heavy computational burden. Moreover, since the inversion of certain rational matrices which necessitates cancellation of poles and zeros is required in the recovery of optimal controller from the resulting optimal closed-loop map, this line of approaches will commonly suffer from numerical difficulties given the finite computational precision inherent in any numerical scheme.

Lately, a new methodology was proposed in [2] to solve the ℓ_1 minimization problem. This approach, referred as the Scaled-Q method, avoids the utilization of zero interpolation to characterize the admissible closed-loop maps. Also it yields the impulse response sequence of the minimizing Youla parameter as the optimal solution. This makes straightforward the task of controller recovery. More noticeably, this approach suggests that introducing a norm bound on the Youla parameter Q in minimization may lead to a well regularized optimization problem. Motivated by this idea and by appealing to result of linear operator theory, solutions to the $\mathcal{H}_2/\ell_1$ problem, the $\ell_1/\mathcal{H}_2$ problem, and the $\ell_1 + \mathcal{H}_2$ combination problem have been obtained([3],[4]). The same idea was also used in [16] to obtain the lower bound for the $\mathcal{H}_2 + \mathcal{H}_\infty$ problem.

A common engineering convention is to state some of the performance requirements in terms of the transient time response of the closed-loop system to a given test signal (such as a step). It has been well recognized that standard single-norm control (ℓ_1, $\mathcal{H}_2$, or $\mathcal{H}_\infty$) cannot handle specifications or constraints on the time response of a closed-loop system exactly. Thus, there exists a need to consider the time response specifications explicitly in the multi-objective problem setup. When the involved objective norm is the ℓ_2 gain of the closed-loop map, the corresponding multi-objective problem with finite horizontal TDC was solved in [27]-[29]. Solutions were obtained in [30] to the case when the template constraints were imposed over an infinite horizon. In [31], the problem of $\mathcal{H}_2$ minimization with constraints on the time-domain response of a closed-loop transfer function was studied. By using weighting functions and absorbing these weights in the generalized plant, the TDC was (approximately) translated into an ℓ_∞ bound on the closed-loop map of interest, and then the problem was solved by solving a sequence of finite-dimensional quadratic programming problems. In [32], an algorithm was proposed to explicitly obtain the state feedback control law to minimize a quadratic performance criterion with TDC on inputs and states. For both the finite horizon and the usual infinite horizon problems, the control laws were shown to be piecewise linear and continuous. In [19] and [33], the ℓ_1/TDC mixed problem was addressed using zero interpolation method.

5 The Constrained Q Method

As illustrated by the suspension control example introduced in Section 1, applied engineering system designs have been calling for more advanced control synthesis theory which can yield optimal controllers satisfying multiple criteria simultaneously. More explicitly, from the viewpoint of applications, there is a need for developing solutions to the multi-objective problem accommodating more than three diverse objectives all together (e.g., $\ell_1/\mathcal{H}_2/\mathcal{H}_\infty$). Towards this goal, let us consider the following (general) multi-objective problem setup: *Consider exogenous inputs w_i and regulated outputs z_i of interest. Let $w_7 = r$ and $z_7 = y$ denote a given scalar reference input and the system time response output respectively. Given constants c_i, $i = 1, \ldots, 6$, and two sequences $\{Lower(k)\}_{k=0}^{\infty}$ and $\{Upper(k)\}_{k=0}^{\infty}$, solve the following problem,*

$$\begin{aligned} &\inf_{\text{K stabilizing}} \quad c_1\|R^1(K)\|_1 + c_2\|\hat{R}^2(K)\|^2_{\mathcal{H}_2} + c_3\|\hat{R}^3(K)\|_{\mathcal{H}_\infty} \\ &\text{subject to} \\ &\qquad \|R^4(K)\|_1 \le c_4 \\ &\qquad \|\hat{R}^5(K)\|^2_{\mathcal{H}_2} \le c_5 \\ &\qquad \|\hat{R}^6(K)\|_{\mathcal{H}_\infty} \le c_6 \\ &\qquad Lower(k) \le y(k) \le Upper(k),\ k = 0, 1, 2, \ldots \end{aligned} \tag{4}$$

where $\{y(k)\}_{k=0}^{\infty}$ denotes the time response of the closed-loop system due to the exogenous reference input r while $w_i = 0$, $i = 1, \ldots, 6$. In the sequel, with a little abuse of notation, we shall use $c_i = \infty$ $(i = 4, 5, 6)$ to denote the case where the corresponding $\ell_1/\mathcal{H}_2/\mathcal{H}_\infty$ norm constraint on R^i $(i = 4, 5, 6)$ is inactive. The problem (4) defined above represents a large class of multiobjective control problems. In particular, it accommodates all the multiobjective problems we have surveyed in the last section (e.g., $\ell_1/\mathcal{H}_2$ corresponds to a special case of problem (4) where $c_2 = c_3 = 0$, $c_4 = c_6 = \infty$, and TDC on y is inactive).

Now based on the Youla parametrization, we define the following auxiliary problem which is closely related to the above problem (assuming $c_i > 0$). The statement of the auxiliary problem is given as follows:

$$\begin{aligned} &\inf_{Q \in \ell_1^{n_u \times n_v}} \quad f(Q) \\ &\text{subject to} \\ &\qquad \|Q\|_1 \le \gamma \\ &\qquad \|R^4(Q)\|_1 \le c_4 \\ &\qquad \|R^5(Q)\|_2^2 \le c_5 \\ &\qquad \|\hat{R}^6(Q)\|_{\mathcal{H}_\infty} \le c_6 \\ &\qquad a_{temp}(k) \le y(k) \le b_{temp}(k),\ k = 0, 1, 2, \ldots. \end{aligned} \tag{5}$$

where $f(Q) := c_1\|R^1(Q)\|_1 + c_2\|R^2(Q)\|_2^2 + c_3\|\hat{R}^3(Q)\|_{\mathcal{H}_\infty}$, $R^i(Q) = H^{ii} - U^i * Q * V^i, i = 1, \ldots, 7$.

Note there is an extra ℓ_1 norm constraint on the Youla parameter Q in the auxiliary problem. It can be shown that if $\hat{U}^2$ and $\hat{V}^2$ have full column and row rank on the unit circle respectively, then γ can be determined a priori such that the constraint $\|Q\|_1 \le \gamma$ is inactive and the problem (4) and (5) have the same optimal costs. It is shown in [34] that converging lower and upper bounds to problem (5) can be computed by solving two sequences of finite dimensional semidefinite programming problems. As mentioned earlier, the above idea of introducing an extra norm bound on the Youla parameter Q to regularize the problem was also adopted in [16] to compute lower bounds to the $\mathcal{H}_2 + \mathcal{H}_\infty$ combination problem.

It is worth mentioning that, in the methodology introduced above, the impulse response of the Youla parameter Q is taken as the optimizing variable. This has similarity to the so called Q-des method introduced in [35]. In the latter method, although upper bounds for the optimal performance of (4) can be guaranteed, there is no analytical result available on the convergence property of the suboptimal cost sequence obtained by increasing supported length of Q. Actually there was no known method existing in the literature that is capable of furnishing a lower bound for problem (4). In our method, the introduction of the regularizing condition $||Q||_1 \leq \gamma$ renders us a well behaved optimization problem enabling to establish the existence of an optimal solution to (5) and converging lower and upper bounds, which is not possessed by the Q-des method.

6 Conclusion

In this paper we have presented a number of newly-developed approaches that provides solutions to the multi-objective nominal performance problems. Also we introduced a new attractive methodology to solve the general multi-objective problem involving all the three commonly used norm objectives ($\ell_1/\mathcal{H}_2/\mathcal{H}_\infty$) and the explicit time-domain constraint specification (TDC). Future Work is needed to implement the proposed methodology into more accessible software.

It should be mentioned that although some preliminary results are available in the literature(see for instance [36], [37], and [38]), the multi-objective robust performance problem still remains in the infant stage. More attention is needed along this direction.

References

[1] B. D. Jager, “Multiobjective suspension control problem,” *Proc. of CDC95,pp.3652-3657*, 1995.

[2] M. H. Khammash, “A New Approach to the Solution of the ℓ_1 Control Problem: The Scaled-Q Method,” *IEEE TAC,Vol.45,No.2,pp.180-187*, 2000.

[3] M. V. Salapaka, M. Khammash, and M. Dahleh, “Solution of MIMO $\mathcal{H}_2/\ell_1$ problem without zero interpolation,” *Proc. of CDC97,pp.1546-1551*, 1997.

[4] M. V. Salapaka, M. H. Khammash, and M. Dahleh, “Solution of MIMO $\mathcal{H}_2/\ell_1$ Problem without Zero Interpolation,” *SIAM J. on Contr. and Optim.,Vol.37,No.6,pp.1865-1873*, 1999.

[5] N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory.* Interscience Publishers, Inc., 1967.

[6] S. P. Boyd and C. H. Barratt, *Linear Controller Design: Limits of Performance.* Prentice Hall, 1991.

[7] N. Elia, *Computational Methods for Multi-Objective Control.* PhD thesis, MIT, Cambridge, MA 02139, 1996.

[8] B. Vroemen and B. d. Jager, “Multiobjective control: An overview,” *Proc. of CDC97,pp.440-445*, 1997.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, 1985.

[10] M. A. Dahleh and I. J. Diaz-Bobillo, *Control of Uncertain Systems: A Linear Programming Approach.* Prentice Hall, 1995.

[11] C. W. Scherer, “Multiobjective $\mathcal{H}_2/\mathcal{H}_\infty$ control,” *IEEE Trans. on Auto. Contr.(TAC),Vol.40,pp.1054-1062*, 1995.

[12] C. W. Scherer, P. Gahinet, and M. Chilali, "Multiobjective output-feedback control via LMI optimization," *IEEE TAC, Vol.42,pp.896-911*, 1997.

[13] H. A. Hindi, B. Hassibi, and S. P. Boyd, "Multiobjective $\mathcal{H}_2/\mathcal{H}_\infty$-Optimal Control via Finite Dimensional Q-Parametrization and Linear Matrix Inequalities," *Proc. of ACC98,pp.3244-3248*, 1998.

[14] I. Masubuchi, A. Ohara, and N. Suda, "LMI-based Controller Synthesis: A Unified Formulation and Solution," *Int. J. Robust and Nonlin. Contr., Vol.8,pp.669-686*, 1998.

[15] C. W. Scherer, "From Mixed to Multi-Objective Control," *Proc. of CDC99,pp.3621-3626*, 1999.

[16] C. W. Scherer, "Lower Bounds in Multi-Objective $\mathcal{H}_2/\mathcal{H}_\infty$ Problems," *Proc. of CDC99,pp.3605-3610*, 1999.

[17] M. C. de Oliveira and J. C. Geromel and J. Bernussou, "An LMI optimization approach to multiobjective controller design for discrete-time systems," *Proc. of CDC99,pp.3611-3616*, 1999.

[18] K. R. Horspool, R. S. Erwin, C. Abdallah, and P. Dorato, "A Randomized approach to the $\mathcal{H}_2/\mathcal{H}_\infty$ Control Problem via Q-parametrization," *Proc. of ACC00,pp.1400-1404*, 2000.

[19] N. Elia and M. A. Dahleh, "Controller Design with Multiple Objectives," *IEEE TAC, Vol.42,No.5,pp.596-613*, 1997.

[20] S. R. Venkatesh and M. A. Dahleh, "Examples of Extreme Cases in ℓ_1 and $\mathcal{H}_\infty$ optimization," *Proc. of CDC93,pp.2188-2193*, 1993.

[21] M. Sznaier and J. Bu, "Mixed $\ell_1/\mathcal{H}_\infty$ Control of MIMO Systems via Convex Optimization," *IEEE TAC, Vol. 43,No.9,pp.1229-1241*, 1998.

[22] H. Rotstein, "Convergence of optimal control problems with an $\mathcal{H}_\infty$-norm constraint," *Automatica, Vol.33,No.3,pp.355-367*, 1997.

[23] H. Rotstein and B. Shklyar, "Optimal Solution and Approximation to the $\ell_1/\mathcal{H}_\infty$ Control Problem," *IEEE TAC, Vol.44,No.6,pp.1240-1243*, 1999.

[24] M. V. Salapaka and M. Dahleh, *Multiple Objective Control Synthesis.* Springer-Verlag(Monograph), 1999.

[25] M. V. Salapaka, M. Dahleh, and P. G. Voulgaris, "Mixed Objective Control Synthesis: Optimal $\ell_1/\mathcal{H}_2$ Control," *SIAM Journal on Control and Optimization, Vol.35No.5,pp.1672-1689*, 1997.

[26] M. V. Salapaka, M. Dahleh, and P. G. Voulgaris, "MIMO Optimal Control Design: The Interplay Between the $\mathcal{H}_2$ and the ℓ_1Norms," *IEEE TAC, Vol.43,No.10,pp.1374-1388*, 1998.

[27] H. Rotstein and A. Sideris, "$\mathcal{H}_\infty$ Optimization with Time-Domain Constraints," *IEEE TAC, Vol.39,No.4,pp.762-779*, 1994.

[28] S. Hosoe, "LMI Approach to an $\mathcal{H}_\infty$-Control Problem with Time-Domain Constraints over a Finite Horizon," *IEEE TAC, Vol. 43,No.8,pp.1128-1132*, 1998.

[29] V. Balakrishnan, "$\mathcal{H}_\infty$-Controller Synthesis with Time-Domain Constraints," *IEEE TAC, Vol. 42,No.8,pp.1179-1186*, 1998.

[30] A. Sideris and H. Rotstein, "Constrained $\mathcal{H}_\infty$ Optimal Control over an Infinite Horizon," *SIAM J. on Contr. and Optim., Vol.35,No.4,pp.1244-1262*, 1997.

[31] M. Sznaier and T. Amishima, "$\mathcal{H}_2$ Control with Time Domain Constraints," *Proc. of ACC98, Vol.5,pp.3229-3233*, 1998.

[32] A. Bemporad, M. Morari, V. Dua, and E. N. Pistikopoulos, "The Explicit Linear Quadratic Regulator for Constraint Systems," *Technical Report AUT99-16, ETH Zürich*, 1999.

[33] Y. Ohta and Y. Toude, "Minimization of the Maximum Peak-to-peak Gain with Time Domain Constraints on Fixed Input Response," *Proc. of CDC99, Phoenix Arizona,pp.1439-1444*, 1999.

[34] X. Qi, M. H. Khammash, and M. V. Salapaka, "Optimal Controller Synthesis with Multiple Objectives," *submitted to ACC 2001.*

[35] S. P. Boyd, V. Balakrishnan, C. H. Barratt, N. M. Khraishi, X. Li, D. G. Meyer, and S. A. Norman, "Approximation Methods of Scalar Mixed $\mathcal{H}_2/\ell_1$ Problems for Discrete-Time Systems," *IEEE TAC, Vol.33,No.13,pp.268-283*, 1988.

[36] M. H. Khammash, M. V. Salapaka, and T. Vanvoorhis, "Synthesis of globally optimal controllers in ℓ_1 using linear relaxation," *Proc. of CDC98,pp.3315-3320*, 1998.

[37] J. Bu and M. Sznaier, "Robust Analysis with Mixed Performance Objectives," *Proc. of CDC97*, 1997.

[38] M. V. Salapaka, M. Dahleh, A. Vicino, and A. Tesi, "Nominal $\mathcal{H}_2$ performance and ℓ_1 robust performance," *Proc. of CDC96,pp.4034-4039*, 1996.

[39] M. V. Salapaka, M. H. Khammash, and M. Dahleh, "Recent Progress in Mixed Objective Control Synthesis with ℓ_1 and $\mathcal{H}_2$ Norms," *Proceedings of the ASME conference*, 1998.

Performance Evaluation of Distributed Control Systems With Reduced Communication*

J. K. Yook, D. M. Tilbury
Dept. of Mechanical Engineering
The University of Michigan
Ann Arbor, MI 48109
{jyook, tilbury}@engin.umich.edu

Abstract

In many distributed control systems, the finite bandwidth of the communication channel is the limiting factor in scaling up a system. To reduce the amount of actual communication needed, estimators at each node can be used. In essence, increased computational costs are traded for decreased communication. By using the estimated output values instead of the true output values in the control algorithm at each sample time, a significant savings in the required bandwidth is achieved. The performance of the system, however, is necessarily degraded from the baseline centralized performance. Nevertheless, the performance degradation of the distributed control system with the proposed estimator can be explicitly analyzed and determined as a function of the threshold value. In this paper, we show how to choose a threshold value to guarantee a desired system performance. The analysis results are compared with simulation on a simple two degree-of-freedom linear system, and experimental results on a nonlinear system are included to show the effectiveness of the proposed framework in a real system.

1 Introduction

When a control system is implemented in a distributed fashion, with multiple processors communicating over a network, the communication delays associated with the network as well as the computation delays associated with the processing time can degrade the system's performance. However, due to the rapid development of the microprocessor, the effects of the computation delay on the overall performance of the system are becoming less significant. On the other hand, the communication delay still has a major effect on the system performance and can be a major problem especially in a system with many nodes and a high communication frequency. To address the communication problem, there are two main directions towards realizing complex MIMO systems. The first uses a decentralized controller to eliminate the communication, and the second uses improved communication algorithms to minimize the communication delay. However, there have been few attempts to reduce the communication required.

In our previous work, we have shown how estimators at each node can be used to minimize the amount of actual communication needed [13]. In essence, increased computational costs are traded for decreased communication. If the estimated value differs

---

*This research was supported in part by the NSF under the grants EEC95-92125 and CMS 99-77179.

from the true value by more than a predetermined threshold, the true value is broadcast to all other nodes in the system. By using the estimated output values instead of the true output values in the control algorithm at each sample time, a significant savings in the required bandwidth is achieved. The performance of the system, however, is necessarily degraded from the baseline centralized performance. Nevertheless, the performance degradation of the distributed control system with the proposed estimator can be explicitly analyzed and determined as a function of the threshold value. In this paper, we show how to choose a threshold value to guarantee a desired system performance. The analysis results are compared with simulation on a simple two degree-of-freedom linear system. In order to reduce the required communication further, a state recovery algorithm to update estimator states is proposed; we demonstrate the validity of the algorithm through simulation. Furthermore, experimental results on a nonlinear system are included to demonstrate the effectiveness of the proposed framework in a real system.

Although to the authors' knowledge, the proposed estimation framework is novel and the performance issues with reduced communication have not been examined previously in a distributed control framework, there are research results in related areas such as dead-reckoning [2, 7], control network effects [1, 6, 10] and time-delays effects [3, 5, 4]. More details can be found in [11, 13].

2 The State Estimator Framework

The proposed state estimator is a centralized MIMO system which contains the nominal plant P with A_o, B_o, C_o (which is used to design the controller) instead of the actual plant Π with A, B, C. On the i^{th} node, the input to the controller is composed of the actual error E_i^* for the output associated with that node, and the estimated errors $\hat{E}_j, j \neq i$ for all other nodes. In addition, the proposed state estimator includes communication logic which compares the actual output of the i^{th} node with the estimated output of the i^{th} node and manages the communication from the i^{th} node to the entire system. When communication delays are negligible, the performance of a distributed control system should be the same as that of a centralized MIMO system. Thus, a MIMO system without network induced delay is chosen as the nominal system to which the performance of the system with the proposed estimator scheme will be compared.

2.1 System State Equations

Consider a discrete time, causal, LTI MIMO system which can be described as follows

$$X(k+1) = A \cdot X(k) + B \cdot U(k) + B \cdot \mathbf{D}(k), \quad Y(k) = C \cdot X(k) \tag{1}$$

with feedback control

$$U(z) = K(z) \cdot E(z) = K(z) \cdot [R(z) - (C \cdot X(z) + \mathbf{N}(z))] \tag{2}$$

where R, $\mathbf{D}$, $\mathbf{N}$, and K are the reference, disturbance, sensor noise, and closed-loop controller, respectively. Then the states of the system can be expressed as follows

$$X(z) = [z \cdot I - \Psi(z)]^{-1} \cdot [B \cdot K(z) \cdot R(z) - B \cdot K(z) \cdot \mathbf{N}(z) + B \cdot \mathbf{D}(z)] \tag{3}$$

where $\Psi(z) = (A - B \cdot K(z) \cdot C)$.

When the system described in Equation (1) and (2) is implemented with the proposed estimator, it can be expressed by the following equations:

$$X^*(k+1) = A \cdot X^*(k) + B \cdot U^*(k) + B \cdot \mathbf{D}(k), \quad Y^*(k) = C \cdot X^*(k) \tag{4}$$

Table 1: Comparison between updating methods.

updating method	assumptions	data used	data updated
output update	$\exists j, s.t.\ Y_i = x_j$	Y_i^*	$\hat{x}_j \leftarrow Y_i^*$
state recovery update	observable	$\tilde{Y}_i = \{Y_i^*, \hat{Y}_j, j \neq i\}$ $\tilde{U}_i = \{U_i^*, \hat{U}_j, j \neq i\}$	est. $X_i^* \leftarrow f(\tilde{Y}_i, \tilde{U}_i)$ $\hat{X}_i \leftarrow$ est. X_i^*
state update	all states measured	X_i^*	$\hat{X}_i \leftarrow X_i^*$

Even though the above equation is identical in form to Equation (1), the controller output U^* differs from U in Equation (1), since U^* is computed using both actual and estimated errors. The differences between the actual and estimated outputs Γ can be described as;

$$\Gamma = (Y^* + \mathrm{N}) - \hat{Y} = (R - \hat{Y}) - (R - (Y^* + \mathrm{N})) = \hat{E} - E^* \tag{5}$$

The actual controller output in the system with the estimator U^* can be expressed as;

$$U^* = K \cdot E^* + K \cdot \Gamma - K_{\text{dia}} \cdot \Gamma = K \cdot E^* + K_{\text{off}} \cdot \Gamma \tag{6}$$

Using Equations (4) and (6), the states of the system with the estimator are;

$$X^*(z) = [zI - \Psi(z)]^{-1} \left[BK(z)R(z) + BK_{\text{off}}(z)\Gamma(z) - BK(z)\mathrm{N}(z) + B\mathrm{D}(z)\right] \tag{7}$$

When Equations (3) and (7) are compared, they only differ by the additional term $BK_{\text{off}}\Gamma$. However, with the communication logic, Γ will be bounded by the threshold value H so that we can treat Γ as a bounded disturbance input to the system.

2.2 State Update Methods

When there is a communication from i^{th} node, the estimators in all nodes are updated to reflect the new information.[1] As summarized in Table 1, either the output or the state can be updated. If the outputs are a subset of the state variables, the output update method can be used. Since only the measured states are updated, the unmeasured states will always differ from their actual values. Furthermore, even at a communication instant, the estimated state set does not reflect the estimator output set. To overcome these limitations, the actual states can be approximately reconstructed or the state update method can be used if all states are measured.

To reconstruct the actual states, we would need all the sensor measurements and control outputs. Communication between all nodes would thus be required, which is undesirable. Therefore, we propose to use a combination of estimated and actual values to approximately recover the states related to the i^{th} node. To determine which states can be reconstructed from the i^{th} output, the observer form can be used.[2] The basic idea of the state recovery algorithm is to break the MIMO system into a set of subsystems $(A_{o_{i,i}}, B_{o_i}, C_{o_{i,i}})$ of the observer form, and assume the estimator states and control outputs of the remote nodes are close to their actual value. Then, using the observability of $(A_{o_{i,i}}, C_{o_{i,i}})$ and considering the remote states as inputs to the i^{th} subsystem, we estimate the states associated with the i^{th} subsystem X_i^* from the output Y_i^*.

---

[1] Note that it is possible to have more than one node sending updated information.

[2] A transformation matrix to observer form of any observable system always exists and can be found in [8], p.232.

The proposed state recovery algorithm can be summarized as follows

$$\begin{aligned} X^*_{i,j}(k) &= \sum_{\xi=1}^{j} \Big[(A_{i,i})_{\xi,j} \cdot X^*_{i,j}(k-j+\xi-1) + B_{i,\xi} \cdot \tilde{U}_i(k-j+\xi-1) \\ &\quad + \Omega_{i,\xi}(k-j+\xi-1) \Big] \end{aligned}$$

where the following notation has been introduced to simplify the expressions:

$$\begin{aligned} \tilde{U}_i &= \begin{bmatrix} \hat{U}_1 & \cdots & \hat{U}_{i-1} & U_i^* & \hat{U}_{i+1} & \cdots & \hat{U}_m \end{bmatrix}^T \\ \Omega_{i,\xi}(k-1) &= \sum_{j=1}^{i-1} (A_{i,j})_{\xi,n_j} \cdot \hat{X}_{j,n_j}(k-1) + \sum_{j=i+1}^{m} (A_{i,j})_{\xi,n_j} \cdot \hat{X}_{j,n_j}(k-1) \end{aligned}$$

Since the state recovery scheme uses both actual and estimated controller outputs $\tilde{U}_i$ along with Ω which is calculated using the estimated states of other remote nodes, the recovered states are not necessarily the actual states. However, we predict that the state recovery algorithm will require less communication than the output update algorithm. Simulation results, presented in Section 4, confirm this prediction.

3 Analysis of the State Estimator System

The purpose of the analysis presented in this section is to examine how the system responds with the proposed estimator scheme. We assume that the original MIMO system is well-designed; meaning that both the actual closed-loop system and its model are exponentially stable.

Theorem 1: Given a finite threshold H, the system with the proposed estimator as defined in Equation (7) is BIBO stable [13].

Theorem 2: Let the performance index $\mathbf{PI}(k) = Y^*(k) - Y(k)$ be the difference between the outputs of the distributed system with the proposed estimator and the nominal centralized system. Then for any desired i^{th} bound $\{\sup_k |\mathbf{PI}_i(k)| \leq \varepsilon_i\}$, there exists a communication threshold value H which guarantees this $\mathbf{PI}_i$ bound is satisfied. Furthermore, this H is independent of noise, disturbances, and the estimator plant model used in the system.

Proof of Theorem 2: The proof is by direct computation. Using the definition of $\mathbf{PI}$ and Equations (3) and (7), we get

$$\mathbf{PI}(z) = Y^*(z) - Y(z) = \Phi_{\mathbf{PI}}(z) \cdot \Gamma(z) \tag{8}$$

where $\Phi_{\mathbf{PI}}(z) = C \cdot [z \cdot I - \Psi(z)]^{-1} \cdot B \cdot K_{\text{off}}(z)$. By Theorem 1, $\Phi_{\mathbf{PI}}(z)$ is an exponentially stable transfer function matrix. Thus, the input-output behavior of Equation (8) can be specified by the unit-pulse response,

$$\mathbf{PI}(k) \;=\; \sum_{\ell=0}^{k-1} G(\ell) \cdot \Gamma(k-\ell), \quad k \geq 1$$

where $G(k) = \mathcal{Z}^{-1}\{\Phi_{\mathbf{PI}}(z)\}$. Then, $|\mathbf{PI}_i|$ satisfies the following inequalities,

$$\begin{aligned} |\mathbf{PI}_i(k)| &= \left| \sum_{\ell=0}^{k-1} G_i(\ell) \cdot \Gamma(k-\ell) \right| \leq \sum_{\ell=0}^{k-1} \left| \sum_{j=1}^{m} G_{i,j}(\ell) \cdot \Gamma_j(k-\ell) \right| \\ &\leq \sum_{\ell=0}^{k-1} \sum_{j=1}^{m} |G_{i,j}(\ell)| \cdot \|\Gamma(k-\ell)\|_\infty = \|G_i\|_1 \cdot \|H\|_\infty \end{aligned} \tag{9}$$

where G_i and $G_{i,j}$ indicate i^{th} row of G and $(i,j)^{th}$ entry of G, respectively. Choose the threshold bound H such that $\|H\|_\infty \leq \frac{\varepsilon_i}{\|G_i\|_1}$. Since $\Phi_{\mathbf{PI}}(z)$ is not a function of $\mathbf{N}$, $\mathbf{D}$, or (A_o, B_o, C_o), the chosen threshold value H is independent of noise, disturbances, and the estimator plant model used in the system.$\diamondsuit$

Remark 1: In Theorem two, we have shown that H such that $\|H\|_\infty \leq \frac{\varepsilon}{\|G_i\|_1}$ is an upper bound on $\max(|\mathbf{PI}_i|)$. In fact, this is the least upper bound.

Remark 2: The existence of H satisfying the bound $\{\sup_k |\mathbf{PI}_i(k)| \leq \varepsilon_i\}$ can be easily seen by the definition of $\mathbf{PI}$. Since the lowest bound $\|\mathbf{PI}(k)\| = 0$ is defined with $H = 0$, for any desired bound ε_i, $\|H\|_\infty = 0$ can be used for the threshold value at least.

Remark 3: To evaluate the output signal in terms of a system norm, information about the input signal is required. In Theorem 2, for example, given $\max(\Gamma_i) \leq H_i$, $\|G_i\|_1$ is needed to get the bound on $|\mathbf{PI}_i|$. Note that for $\|\mathbf{PI}\|_\infty \leq \varepsilon$, we can choose H such that $\|H\|_\infty \leq \frac{\varepsilon}{\max_i \|G_i\|_1}$. However, if Γ could be approximated with a bounded set of sinusoidal functions, singular value analysis could be applied to get an identical bound on $\|\mathbf{PI}\|_\infty$ [14, p.104].

It is true that Γ will be influenced by noise, disturbances, the estimator plant model used; however, $\|\Gamma\|$ is limited by H. Thus, when a system is realized with the proposed estimator, the difference in performance between a distributed and centralized implementation is bounded. Note that the actual system parameters (A, B, C) are used in Equation (8) and (9) to compute the threshold H in order to guarantee the performance of the system with the estimator. Although we cannot know the actual system parameters in a real system, robust control theory such as μ-synthesis can be used to relate modeling uncertainty to $\|\Phi_{\mathbf{PI}}(z)\|_\infty$. Since we use the threshold value H on $(Y^* + \mathbf{N}) - \hat{Y}$, a rule of thumb is to choose $\|H\|_\infty$ close to ε. This is due to the fact that $\Phi_{\mathbf{PI}}$ is very close to the closed loop transfer function of the system except for K_{off} instead of K.

The communication frequency depends on the behavior of $\Gamma = Y^* - \hat{Y}$. In other words, when $|\Gamma_i| \geq H_i$ is true, a communication from the i^{th} subsystem occurs and Γ_i becomes a discontinuous variable bounded by the threshold values H_i. That is, Γ_i is a piecewise continuous variable with the discontinuities occurring at the instants of communication. In [13], it is described how Γ can be expressed by

$$\Gamma \;\approx\; \Phi_S \cdot [\Phi_R \cdot R + \mathbf{N}] \tag{10}$$

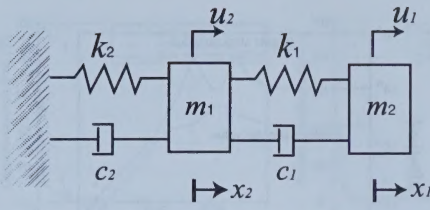
where $\Phi_S(z) = [I - C[zI - \Psi(z)]^{-1} B K_{\text{off}}(z)]^{-1}$ and $\Phi_R(z) = [C[zI - \Psi(z)]^{-1} B - C_o[zI - \Psi_o(z)]^{-1} B_o] K(z)$. For a reasonably well-designed controller, the convergence of Γ depends only on the two transfer matrices Φ_S and Φ_R.

4 Simulation and Experimental Results

4.1 Simulation and Analysis: a Mass-Spring-Damper System

A two degree of freedom mass-spring-damper system is shown in Figure 1a. The dynamical system can be described by the following differential equations:

$$\begin{bmatrix} \dot{x}_1 \\ \ddot{x}_1 \\ \dot{x}_2 \\ \ddot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_1}{m_1} & -\frac{b_1}{m_1} & \frac{k_1}{m_1} & \frac{b_1}{m_1} \\ 0 & 0 & 0 & 1 \\ \frac{k_1}{m_2} & \frac{b_1}{m_2} & -\frac{k_1+k_2}{m_2} & -\frac{b_1+b_2}{m_2} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \dot{x}_1 \\ x_2 \\ \dot{x}_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \frac{1}{m_1} & 0 \\ 0 & 0 \\ 0 & \frac{1}{m_2} \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{11}$$



a.

Variable assignments
to the proposed framework

X_1^* subsystem	X_2^* subsystem
$X_1^* = [x_1, \dot{x}_1]^T$	$X_2^* = [x_2, \dot{x}_2]^T$
$Y_1^* = x_1$	$Y_2^* = x_2$
$U_1^* = u_1$	$U_2^* = u_2$

b.

Figure 1: A mass-spring-damper system

Define $\Pi(s)$ as the transfer matrix from (u_1, u_2) to (x_1, x_2). Let the actual and nominal parameters be as follows

$$\begin{array}{lllllll} A, B, C & \text{from } k_1 = 1.1 & k_2 = 4.4 & b_1 = 0.22 & b_2 = 0.11 & m_1 = 1.1 & m_2 = 2.2 \\ A_o, B_o, C_o & \text{from } k_{1_o} = 1 & k_{2_o} = 4 & b_{1_o} = 0.2 & b_{2_o} = 0.1 & m_{1_o} = 1 & m_{2_o} = 2 \end{array}$$

where the units of k, b, m are N/m, Nm/sec, and Kg, respectively. A controller K is designed to have a integral control action for both reference inputs:

$$K = \begin{bmatrix} \frac{40s^3+145.6s^2+178.9s+71.23}{0.1s^4+2.46s^3+19.55s^2+48.74s} & \frac{-100s^3-520s^2-896s-499.2}{s^5+24.15s^4+175.1s^3+85.35s^2+422.5s} \\ \frac{-40s^3-145.6s^2-178.9s-71.23}{0.5s^5+9.9s^4+51.2s^3+19.55s^2+48.74s} & \frac{100s^3+520s^2+896s+499.2}{0.1s^4+2.9s^3+28.9s^2+84.5s} \end{bmatrix} \tag{12}$$

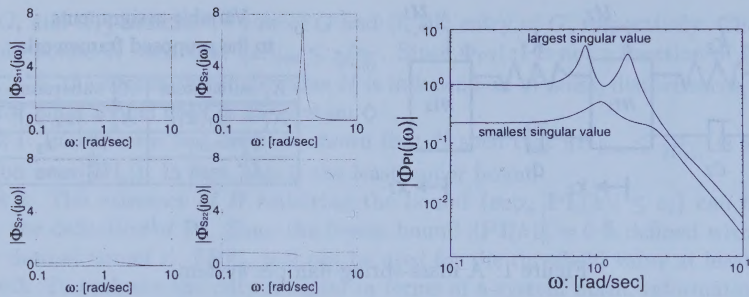
When this system is implemented in a distributed control architecture with the proposed estimator, Figure 1b summarizes the variable assignments.

To analyze this system, the inverse of a MIMO system must be found. Since the inverse of a strictly proper system or an improper system cannot be found, we have added a small additional feed-through term to the strictly proper system, and added some high-frequency dynamics to the improper system in order to get an approximate inverse [9]. With these assumptions, the analysis can be summarized as follows.

- The bode diagram of the transfer matrix Φ_S is shown in Figure 2a. Although Φ_S is stable, it is lightly damped, especially for the second input. Therefore, we expect that Γ_2 will take longer to settle than Γ_1.
- Φ_R is stable and has zero DC gain due to the integral action in K. Since Φ_S and Φ_R are both stable, the effects of the step reference input on Γ will decay to zero.
- Φ_N and Φ_D behave like sensitivity and complementary sensitivity transfer functions, respectively. Both cut-off frequencies are around 1 rad/sec. Therefore, we can use Equation 10 to predict the behavior of Γ for the disturbances and noises bounded by the cut-off frequency, 1 rad/sec.
- $\max_i \|G_i\|_1 = \|G_1\|_1 \approx 9.34$. Therefore, for any performance bound $\|\mathbf{PI}\|_\infty \leq \varepsilon$, H such that $\|H\|_\infty \leq \frac{\varepsilon}{9.34}$ guarantees this performance bound in presence of any noise, disturbances, and the estimator plant model used in the system.
- The singular values of $\Phi_{\mathbf{PI}}$ in Equation (8) are plotted in Figure 2b. The largest singular value is 5.14 at 0.85 rad/sec. If all Γ_i could be approximated with by sinusoidal functions[3], the maximum performance degradation of this system is bounded by the maximum threshold value times 5.14, the largest singular value of $\Phi_{\mathbf{PI}}$.

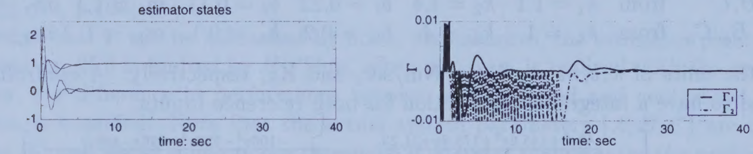
---

[3] This is very strong assumption, and rarely the case. However, this assumption allows us to use μ-syntheses for given modeling error information.

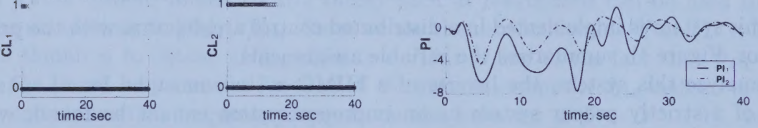


a. Bode plot of Φ_S b. Singular values of $\Phi_{\mathbf{PI}}$

Figure 2: Transfer matrix analysis



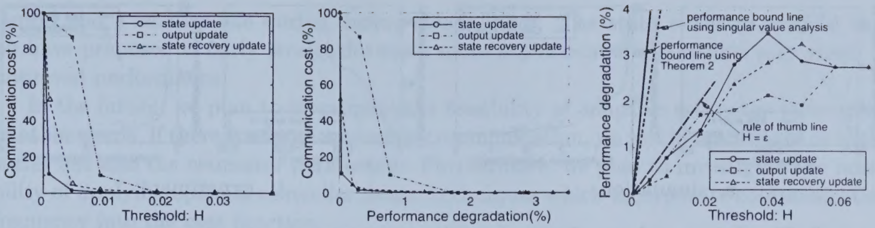
a. Estimator states: $x_1, \dot{x}_1, x_2, \dot{x}_2$ b. $\Gamma = Y^* - \hat{Y}$



c. comm. logic value on output 1 & 2 d. $\mathbf{PI} = Y - Y^*$

Figure 3: A simulation using distributed control with reduced communication and the state recovery algorithm with a threshold value of $H = [0.01, 0.01]^T$.

For the simulation, we set the sampling time $T_s = 0.01$ sec., disturbances $\mathbf{D_1} = \mathbf{D_2} = 0.1\sin(10t)$, and use white noise with $|\mathbf{N_1}|_\infty = |\mathbf{N_2}|_\infty = 0.001$. A simulation result using the state recovery algorithm with threshold $H = [0.01, 0.01]^T$ is shown in Figure 3. In Figure 3c, 1's indicate when a communication occurs from each subsystem. As expected, Γ_2 takes longer to settle than Γ_1. Summaries of the simulations are plotted in the Figure 4. We define the communication cost $\frac{\sum \mathbf{CL}_1 + \sum \mathbf{CL}_2}{2T_{final}/T_s}$ and the performance cost as $\max(\max(|\frac{\mathbf{PI}_1}{R_1}|), \max(|\frac{\mathbf{PI}_2}{R_2}|))$, where $\mathbf{CL}$ is communication logic value at each sampling instant as shown in Figure 3c, $\mathbf{PI}$ is the difference between Y and Y^* as shown in Figure 3e, and R is the reference input. For each threshold value, we examine three updating methods: output update, state recovery update, and state update (assuming all states are measured). In all cases, increasing the threshold values always results in decreased communication. At $H = [0.06, 0.06]^T$, no communication is required within the system. The state update method requires the least communication, followed by the state recovery update method and then the output update method, as expected. We expected that the performance cost would be proportional to the threshold values. However, as shown in Figure 4b, the performance cost has local minima and maxima.



a. b. c.

Figure 4: Summary of the simulation results for the mass-spring-damper system.

This can be explained by the singular values of $\Phi_{\mathbf{PI}}$ shown in Figure 2b. $|\mathbf{PI}|$ depends not only on the magnitude of Γ but also on its frequency content, and different threshold values will result in difference frequency contents in Γ. Due to this phenomenon, the relation between threshold value and the performance cost of each update scheme does not exhibit a well-defined trend. In general, however, increasing the threshold values increases performance degradation until the threshold values reach the point where no communication is required. Nevertheless, the performance degradation is bounded by $||G_1||_1 = 9.34$ times the threshold values even if no communication is required.

Figure 4 clearly shows that the communication required can be significantly reduced with only a minimal impact on the performance of the system. For example, using the estimator with the state recovery algorithm, 1% performance degradation can be tolerated with only 2% of communication cost.[4] More importantly, after the transient response has settled, communication no longer occurs since Γ converges to 0. If this system were realized by a conventional distributed control framework, then the minimum bandwidth of the network should be at least the sampling frequency times the amount of data to send each period. However, with the estimated framework, the bandwidth requirement of the network depends on the threshold value, and is clearly lower than the conventional distributed control framework.

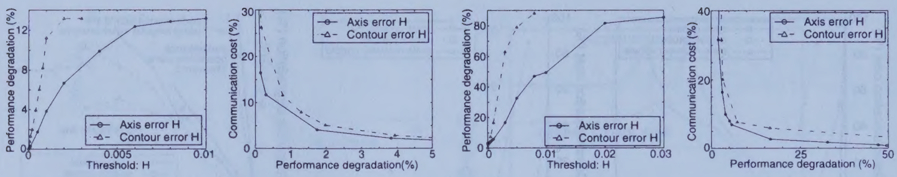
4.2 Simulation and Experiment: a Two Axis System

The experiment we consider is a two-axis table, such as might be used in a milling machine for contouring applications. Although each axis is controlled independently and can be modeled as a linear system, the contour error performance criteria and the cross-coupled controller algorithm used are nonlinear and result in strong coupling between the two axes (detailed descriptions can be found in [13]).

The performance degradation is defined as $(\sum |\varepsilon^* - \varepsilon| / \sum \varepsilon)$, where ε^* and ε denote the contour error of the distributed system with the estimator and contour error of the standard distributed system, respectively. In order to realize instantaneous communication which was one of our assumptions, one computer was used for the experiment emulating instantaneous communication. Due to the limitations of our simple model, which does not include non-linearities such as friction or backlash, we cannot compare the actual values of simulation and experiment directly. Nevertheless, the general trend of the experimental results shown in Figure 5b is similar to that seen in the simulation

---

[4] The relationship between communication cost and performance degradation depends on the modeling error. This issue is addressed in two-axis simulation section.



a. simulation b. experiment

Figure 5: Simulation and experimental results comparing the axis error and the contour error as threshold for the two axis system. For simulation, 10% gain error was assumed. In experiment, there was typically a 2 % performance difference between experiment runs. The proposed state recovery algorithm was used to update the estimator states.

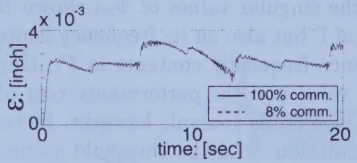


Figure 6: Experimental results for the two axis system. The proposed state recovery algorithm was used to update the estimator states.

results of Figure 5a. In order to achieve within 4% performance degradation, 8% communication cost was required. Figure 6 compares the contour error of 8% and 100% communication, and clearly shows the effectiveness of our framework. These indicate a huge amount of bandwidth savings without significantly reducing performance. Furthermore, even if no communication were required, the contour error would be less than twice the nominal value. In other words, this framework can be used not only to increase network bandwidth but also as an safety feature in case of communication failure.

5 Conclusions and Future Work

As the scale of a distributed control system and its communication frequency increase, the performance of the system can suffer significantly due to network-induced delays. Using the estimator framework, we can decrease the communication requirements of the system, and multivariable controller design and analysis techniques can be employed to improve the performance of the system over a decentralized strategy. Through careful analysis, the performance degradation of the distributed control system with the proposed estimator is explicitly analyzed and determined as a function of the threshold value, and we show how to choose a threshold value to guarantee a desired system performance. The effectiveness of our framework has been validated by simulations and experiments.

The basic estimator framework proposed in this paper lends itself to many different variants. Although we have used a simple absolute value threshold to trigger inter-node communication, different thresholds could be used such as an integration of the differences (ℓ_1 or ℓ_2 norm) or a threshold on the state error rather than the output error. In the case when the transient performance is not as important as the steady-state behavior, time-varying thresholds could be used, with a larger threshold value during the transient period and a smaller one during during steady-state. The state recovery method that we have proposed is fairly straightforward; more sophisticated algorithms may result in improved performance.

In the future, we plan to investigate the feasibility of adaptive estimator techniques. In other words, if there is a large amount of communication, we would update not only the states but also the estimator parameters. Furthermore, we plan to investigate the possibility of applying optimal controller design techniques which incorporate communication frequency into the cost function.

This paper presented a distributed implementation of a centralized control design. On a broader scope, control design issues for distributed control systems remain largely unexplored. To address fault tolerant issues, for example, a decentralized (diagonal) controller may be used to guarantee stability in the system even in the case of total communication failure. The off-diagonal terms in the controller may then be used to improve the performance of the system. Other control architectures and goals will depend on the ever-growing application demands for distributed control systems.

References

[1] K. Arvind, K. Ramamritham, and J. Stankovic. A local area network architecture for communication in distributed real-time systems. *JRTS*, 3(2):115–147, May 1991.

[2] E. Bergund and D. Cheriton. Amaxe: a multiplayer computer game. *IEEE Software*, 2(1), May 1985.

[3] H. Chan and Ü. Özgüner. Closed-loop control of systems over a communication network with queues. *Int. J. Control*, 62:493–510, 1995.

[4] G. F. Franklin, J. D. Powell, and A. Emani-Naeini. *Feedback Control of Dynamic Systems.* Addison-Wesley, Massachusetts, third edition, 1994.

[5] Y. Halevi and A. Ray. Performance analysis of integrated communication and control system networks. *ASME JDSMC*, 112:365–371, September 1990.

[6] F.-L. Lian, J. M. Moyne, and D. M. Tilbury. Performance evaluation of control networks for manufacturing systems. In *Proc. ASME-IMECE, DSCD*, 1999.

[7] A. Pope. The simnet network and protocols. *Technical Report 7102, BBN Systems and Technologies*, July 1989.

[8] W. J. Rugh. *Linear System Theory.* Prentice Hall, New Jersey, 1996.

[9] S. Skogestad and I. Postlethwaite. *Multivariable Feedback Control, Analysis and Design.* John Wiley and Sons Ltd., 1996.

[10] G. C. Walsh, H. Ye, and L. Bushnell. Stability analysis of networked control systems. In *Proc. ACC*, pages 2876–2880, June 1999.

[11] J. Yook, D. Tilbury, and N. Soparkar. A formal characterization of the effects of the time delay on performance in distributed control systems. *Int. J. Control.* To appeare.

[12] J. Yook, D. Tilbury, and N. Soparkar. A design methodolgy for distributed control systems to optimize performance in the presence of time delays. In *Proc. ACC*, 2000.

[13] J. Yook, D. Tilbury, and N. Soparkar. Trading computation for bandwidth: State estimators for reduced communication in distributed control systems. In *Proc. JUSFA*, 2000.

[14] K. Zhou, J. C. Doyle, and Keith Glover. *Robust and Optimal Control.* Prentice Hall, 1996.

Control under Communication Constraints

Sekhar Tatikonda and Sanjoy Mitter

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Increasingly many control systems are employing multiple sensors and actuators that are geographically distributed. There are many issues of coordination and communication that must be dealt with. It becomes important then to determine what signals the sensors should transmit to the controller and what communication channel rates are required to achieve a specified performance.

In this paper we formulate a control problem with a communication channel connecting the plant to the controller. This channel can transmit R bits noiselessly per time step. Our task is to design the encoder and decoder along with the controller to achieve some control objective. We examine traditional control properties, e.g. observability, stability. For a given control objective we provide upper and lower bounds on the rate required to achieve it. In many cases these bounds are tight. In doing so we characterize the "information complexity" of different control objectives.

1 Problem Setup

We consider the following linear time-invariant system:

$$X_0 \in \Lambda_0, \quad X_{t+1} = AX_t + BU_t, \ \forall t \geq 0 \tag{1}$$

where $\{X_t\}$ is a $I\!R^d$-valued state process and $\{U_t\}$ is a $I\!R^m$-valued control process. Where $A \in I\!R^{d \times d}$ and $B \in I\!R^{d \times m}$. The initial position is $X_0 \in \Lambda_0$ where $\Lambda_0 \subseteq I\!R^d$. See Figure 1. We consider a noiseless digital channel that can transmit at each time step one of 2^R symbols $\sigma \in \Sigma$.

The control problems we look at involve the design of the encoder, decoder, and controller. For each component we must specify what information is available to it at each time step. This is often called the information pattern. [Wit]

Encoder: The encoder at time t is a map $\mathcal{E}_t : I\!R^{l(t+1)} \times \Sigma^t \times I\!R^{mt} \to \Sigma$ that takes $(X^t, \sigma^{t-1}, U^{t-1}) \mapsto \sigma_t$. Note that we may restrict the inputs to the encoder. For example the encoder may not have access to the past controls. We will be more specific when we discuss the different encoder/decoder set ups in section 4.

Decoder: Let $\Omega_t \subset I\!R^d$. The decoder at time t is a map $\mathcal{D}_t : \Sigma^{t+1} \times I\!R^{mt} \to 2^{I\!R^d}$ that takes $(\sigma^t, U^{t-1}) \mapsto \Omega_t$. The output of the decoder is a measurable set representing the uncertainty in the state estimate.

Controller: The controller at time t is a map $\mathcal{C}_t : 2^{I\!R^d} \to I\!R^m$ that takes $\Omega_t \mapsto U_t$.

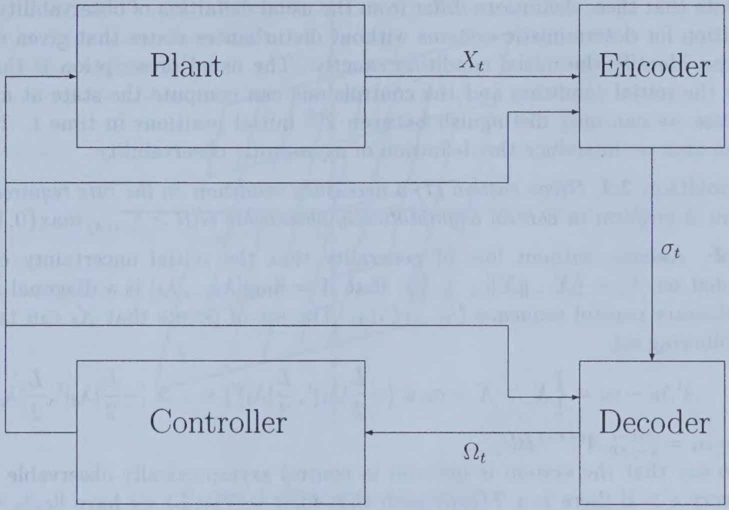


Figure 1: System

2 Lower Bounds that are Independent of the Information Pattern

In this section we provide lower bounds on the rate required to achieve the different control objectives. These lower bounds will be "universal" in the sense that they hold independently of the actual encoder, decoder, and controller used. That is they hold independently of the information pattern chosen. One should note the analogy with Fano's inequality used in converse theorems in information theory.

We will show in section 4 that there exists an information patterns such that an encoder, decoder, and controller exist for which these lower bounds can be achieved. Thus the lower bounds are tight.

Observability

Definition 2.1 *Let the* error *be* $e_t = X_t - \hat{X}_t$ *where* $\hat{X}_t$ *is the state estimate.*

Definition 2.2 *System (1) is* asymptotically observable *if there exists a control sequence* $\{u_t\}$ *and an encoder and decoder such that*

1 *Stability:* $\forall\ \epsilon > 0\ \ \exists\ \delta(\epsilon)$ *such that* $\|X_0\|_2 \leq \delta(\epsilon)$ *implies* $\|e_t\|_2 \leq \epsilon\ \forall t \geq 0$.

2 *Uniform attractivity:* $\forall\ \epsilon > 0,\ \forall\ \delta > 0\ \ \exists\ T(\epsilon, \delta)$ *such that* $\|X_0\|_2 \leq \delta$ *implies* $\|e_t\|_2 \leq \epsilon\ \ t \geq T$.

Definition 2.3 *System (1) is* uniform in control asymptotically observable *if there exists an encoder and decoder independent of the control sequence applied such that the system is asymptotically observable.*

Note that these definitions differ from the usual definition of observability. The usual definition for deterministic systems without disturbances states that given enough time one can identify the initial condition exactly. The usual prescription is that once you know the initial condition and the controls one can compute the state at any time. In our case we can only distinguish between 2^{tR} initial positions in time t. It is for this reason that we introduce the definition of asymptotic observability.

Proposition 2.1 *Given system (1) a necessary condition on the rate required so that the system is uniform in control asymptotically observable is* $R \geq \sum_{\lambda(A)} \max\{0, \log|\lambda(A)|\}$.

Proof: Assume without loss of generality that the initial uncertainty contains the bounded set $\Lambda_0 = \{X : \|X\|_\infty \leq \frac{L}{2}\}$, that $A = \text{diag}[\lambda_1, .., \lambda_d]$ is a diagonal matrix. Fix an arbitrary control sequence $U_0, ..., U_{t-1}$. The set of points that X_t can take contains the following set

$$A^t\Lambda_0 - \alpha_t = \left\{X \ : \ X - \alpha_t \in [-\frac{L}{2}|\lambda_1|^t, \frac{L}{2}|\lambda_1|^t] \times ... \times [-\frac{L}{2}|\lambda_d|^t, \frac{L}{2}|\lambda_d|^t]\right\}$$

where $\alpha_t = \sum_{j=0}^{t-1} A^{t-1-j}BU_j$.

To say that the system is uniform in control asymptotically observable means that for every $\epsilon > 0$ there is a $T(\epsilon, L)$ such that for $t \geq T(\epsilon, L)$ we have $\|e_t\|_2 \leq \epsilon \ \ \forall X_0 \in \Lambda_0, \ \ \forall \ \{U_t\}$.

A lower bound on the rate can be computed by counting the number of regions it takes to cover $A^t\Lambda_0 - \alpha_t$ by regions of diameter less than 2ϵ. If the diameter of the uncertainty set Ω_t goes to zero then the volume of Ω_t goes to zero. The converse, though, is not necessarily true. However the converse is true if we restrict ourselves to computing the volume of the projection of the uncertainty set onto the unstable subspace. The diameter of the uncertainty set goes to zero if and only if the volume of the uncertainty set projected on the unstable subspace goes to zero.

Let $\lambda_1, ..., \lambda_u$ be the unstable eigenvalues. Given a region with diameter 2ϵ project it onto the unstable subspace. The largest volume this projected region can have is that given by a u-dimensional sphere of diameter 2ϵ. Specifically the projected volume is $\leq K_u\epsilon^u$ where K_u is the constant in the formula for the volume of a sphere.

Thus to cover $A^n\Lambda_0 - \alpha_t$ by regions of diameter 2ϵ we require at least

$$R \geq \frac{1}{t+1}\log\frac{L^u|\lambda_1...\lambda_u|^t}{K_u\epsilon^u} = \sum_{\lambda(A)} \max\{0, \log\lambda(A)\} + \frac{u}{t+1}\log\frac{L}{\epsilon} - \frac{1}{t+1}\log|\lambda_1...\lambda_u|K_u$$

the second term is positive and for t large the third term becomes negligible thus $R \geq \sum_{\lambda(A)} \max\{0, \log|\lambda(A)|\}$. □

Stability

Definition 2.4 *System (1) is* asymptotically stabilizable *if there exists an encoder, decoder, and controller such that*

1 Stability: $\forall \ \epsilon > 0 \ \ \exists \ \delta(\epsilon)$ *such that* $\|X_0\|_2 \leq \delta(\epsilon)$ *implies* $\|X_t\|_2 \leq \epsilon \ \forall t \geq 0$.

2 Uniform attractivity: $\forall \ \epsilon > 0, \ \delta > 0 \ \ \exists \ T(\epsilon, \delta)$ *such that* $\|X_0\|_2 \leq \delta$ *implies* $\|X_t\|_2 \leq \epsilon \ \ \forall t \geq T$.

Proposition 2.2 *We are given system (3.1). A necessary condition on the rate for asymptotic stability is* $R > \sum_{\lambda(A)} \max\{0, \log|\lambda(A)|\}$.

Proof: The proof uses a covering argument similar to the one used in proposition 2.1 See [Tat] for full details. □

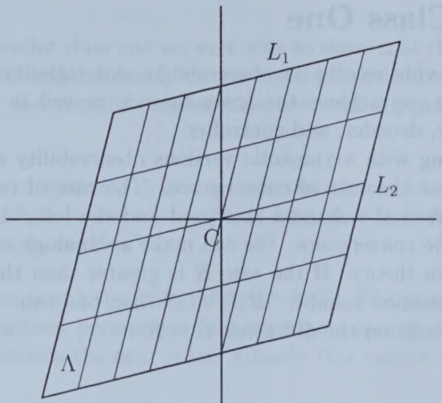


Figure 2: Primitive Quantizer

3 Encoder Classes and Primitive Quantizers

Recall that the encoder at time t is a map $\mathcal{E}_t$ that takes $(X^t, \sigma^{t-1}, U^{t-1}) \mapsto \sigma_t$. In this case the encoder knows the past states, past channel symbols, and past controls. It may seem unreasonable to allow the encoder access to the past controls. For example the encoder may be geographically separated from the plant. We distinguish two encoder structures.
Encoder class 1: In this class the encoder is a map, $\mathcal{E}_t$, that takes $(X^t, \sigma^{t-1}, U^{t-1}) \mapsto \sigma_t$.
Encoder class 2: In this class the encoder is a map, $\mathcal{E}_t$, that takes $(X^t, \sigma^{t-1}) \mapsto \sigma_t$.

Definition 3.1 *A* primitive quantizer *is a four tuple $(C, \underline{R}, \underline{L}, \Phi)$ with $C \in \mathbb{R}^d$ representing the centroid, $\underline{R} = (R_1, ..., R_d) \in \mathbb{R}^{d,+}$ representing the rate vector, $\underline{L} = (L_1, ..., L_d) \in \mathbb{R}^{d,+}$ representing the side-lengths of the dynamic range, and Φ an invertible matrix representing a coordinate transformation. This quantizer partitions the region*

$$\Lambda = \left\{ X \in \mathbb{R}^d \;:\; \Phi(X - C) \in \{[-\frac{L_1}{2}, \frac{L_1}{2}] \times ... \times [-\frac{L_d}{2}, \frac{L_d}{2}]\} \right\},$$

into boxes with side lengths $\frac{L_i}{2^{R_i}}$. Each of the 2^R boxes is represented by an element $\sigma \in \Sigma$. Upon observing X the $(C, \underline{R}, \underline{L}, \Phi)$-quantizer subtracts off C, applies the coordinate transform Φ, determines which box it falls into, and then transmits the σ representing that box. If X falls outside the region Λ then the quantizer transmits a special symbol representing an overflow. Thus we have $2^R + 1$ symbols. The set Λ is called the dynamic range *of the quantizer.*

Figure 2 shows a two-dimensional primitive quantizer with $R_1 = 3$ and $R_2 = 2$. For both encoders 1 and 2 the encoding map $\mathcal{E}_t$ based on its information selects a $(C, \underline{R}, \underline{L}, \Phi)$-quantizer. Upon observing X_t it computes the appropriate σ_t and transmits it across the channel.

4 Encoder Class One

In this section we provide results on observability and stability for encoder class one. We will show that we can achieve the lower bounds proved in section 2 by explicitly describing the encoder, decoder, and controller.

Since we are dealing with asymptotic versions observability and controllability it is of interest to determine the rate of convergence. The rate of convergence will depend on the difference between the channel rate used and the lower bound. The larger this difference the faster the convergence. We can make an analogy with the channel coding theorem in information theory. If the rate R is greater than the capacity C then one cannot transmit information reliably. If $R < C$ then one can. Furthermore the error exponent roughly depends on the difference $C - R$.

Observability

Proposition 4.1 *Given system (1) and encoder in encoder class 1 a sufficient condition on the rate for uniform in control asymptotic observability is* $R > \sum_{\lambda(A)} \max\{0, \log|\lambda(A)|\}$.

Proof: We sketch the proof. For full details see [Tat]. Because the encoder is in encoder class 1 it observes the control signals. Thus the encoder can compute the innovation. This is the difference between the estimate of the state based on the decoder's information and the true state. The encoder will apply the primitive quantizer, with parameters suitably chosen, so that the magnitude of the innovation over time decreases to zero. Note that the uncertainty in the state grows like the eigenvalues of the matrix A. Thus we need to transmit at a rate faster than the growth. But it need only be arbitrarily faster. Thus we get the sufficient condition $R > \sum_{\lambda(A)} \max\{0, \log|\lambda(A)|\}$. □

Stability

For encoder class one we can combine the properties of asymptotic observability and full state feedback stability to get output feedback stability. Assume that the pair (A, B) are stabilizable.

Proposition 4.2 *Given system (1) with encoder in encoder class 1. The rate* $R > \sum_{\lambda(A)} \max\{0, \log|\lambda(A)|\}$ *is a sufficient condition for asymptotic stability.*

Proof: Let K be a stabilizing controller, i.e. $A + BK$ is stable. Apply the certainty equivalent controller $U_t = K\hat{X}_t$ where $\hat{X}_t$ is the centroid of the decoder output Ω_t. We need to show that under this controller the system is stable. Let $e_t = X_t - \hat{X}_t$. Then

$$X_t = (A + BK)^t X_0 - \sum_{j=0}^{t-1} (A + BK)^{t-1-j} BKe_j$$

By proposition 4.1 the system is asymptotically observable under any control sequence. Thus e_t is going to zero. Since $A + BK$ is stable the first addend in the above equation goes to zero. It can be shown the second does also. Hence $\lim_{t\to\infty} X_t = 0$. □

5 Encoder Class Two

In this section we examine observability and stability for encoders in encoder class two. Recall that these encoders do not have access to the past controls.

Observability

For encoders in encoder class one we were able to show that there exist encoders such that observability holds independently of the control signals chosen. Here we will show that observability for encoders in encoder class two will depend on the control signals chosen.

Proposition 5.1 *Given system (1) and encoder in encoder class two. If we allow for arbitrary control signals then there is no finite rate such that the system is uniform in control asymptotically observable.*

Proof: Assume that at time t the uncertainty set is Ω_t. The question then becomes how do we update the quantizer parameters in a way independent of the control such that the dynamic range contains the next state. Clearly this cannot be done uniformly over all control signals. □

Note that in the case of encoder class one we could subtract out the effect of the control and thus could allow for arbitrarily large control signals. In conclusion, encoders in encoder class two cannot achieve asymptotic observability without limiting the controls.

These observations may appear to be bad news. However in most cases we are interested in observability when the controls being applied are in the loop. The statements so far have been about observability uniform over some set of controls. We will show that we can achieve asymptotic observability if we consider the control objective of stability.

Stability

Here we treat stability under encoder class two. Assume (A, B) are controllable.

Proposition 5.2 *Given system (1) with encoder in encoder class two. Then there exists a finite rate such that the system can be made asymptotically stable.*

Proof: We sketch the proof. See [Tat] for full details. The idea here is that the encoder does not know the control signals being applied. But it does know that the control objective is stability. Thus we can provide the encoder with a schedule. This schedule specifies a region where the state will be. Since the objective is stability we know that these regions will shrink. The encoder then at every time step just applies a primitive quantizer tuned to the scheduled region. It can be shown that there is a finite rate such that the size of the regions specified by the schedule shrinks to zero. □

We conclude this section with some general comments on stability under rate constraints. Our schemes involve keeping track of the uncertainty. For encoder class one this involves keeping track of the state estimation error. For encoder class two this involves keeping track of the state but not the error. It is natural to work with these regions. The definition of stability, in particular uniform attractivity, requires their existence. By computing these regions and explicitly stating how they evolve we are able to determine what rate vectors and primitive quantizers are needed.

6 Systems with Multiple Sensors

In this section we examine a particular control problem with multiple sensors. We first formulate the problem. Consider the linear system with M distributed observations and one controller:

$$X_{t+1} = AX_t + BU_t, \quad Y(i)_t = C_i X_t \quad i = 1, \dots M \tag{2}$$

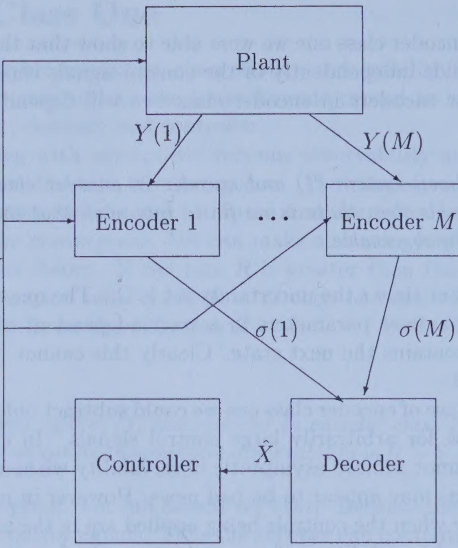


Figure 3: Distributed Sensor System

where $A \in I\!R^{d\times d}, B \in I\!R^{d\times m}, C_i \in I\!R^{l_i\times d}$, and $X_o \in \Lambda_0$. See figure 3.

Let $C \triangleq [C_1', ..., C_M']'$. Assume that the pair (A, C) is detectable and the pair (A, B) is stabilizable. Though (A, C) is detectable it is not generally the case that (A, C_i) will be detectable. Define $\mathcal{N}_i$ to be the unobservable subspace associated with (A, C_i). Define $\mathcal{O}_i$ to be the subspace orthogonal to $\mathcal{N}_i$. It can be shown that $\mathcal{N}_i$ is indeed a subspace and furthermore that it is A-invariant. [Son] By assumption we have $\cap_{i=1}^{M}\mathcal{N}_i = \emptyset$. Also any $x \in \mathcal{O}_i$ can be observed by encoder i.

Associate with each i the set $\Lambda_i = \{\lambda(A) :$ those eigenvalues of A corresponding to the subspace $\mathcal{O}_i\}$. We have $\cup_{i=1}^{M}\Lambda_i = \{\lambda(A)\}$. In general the Λ_i will not be disjoint. Because they are not disjoint we have freedom in determining what each encoder sends to the decoder.

We will show that this freedom can be captured by the following "rate region." For each encoder $\mathcal{E}_i$ define its rate vector to be $\underline{R}_i = (R_{i,1}, ..., R_{i,\dim \mathcal{O}_i})$. (Where $\underline{R}_i$ is the rate vector used in the primitive quantizer by $\mathcal{E}_i$.)

Definition 6.1 *For a given matrix A and observation matrices $C_1, ..., C_M$ define*

$$\mathcal{R} = \left\{ (\underline{R}_1, ..., \underline{R}_M) : \sum_{i:\lambda(A)\in\Lambda_i} R_{i,j_{\lambda(A)}} \geq \max\{0, \log|\lambda(A)|\} \quad \forall \lambda(A) \right\}$$

where $j_{\lambda(A)}$ represents the index of the rate component associated with that eigenvalue.

Proposition 6.1 *Given system (2) a necessary condition on the rate so that the system is uniform in control asymptotically observable is that $(\underline{R}_1, ..., \underline{R}_M) \in \mathcal{R}$.*

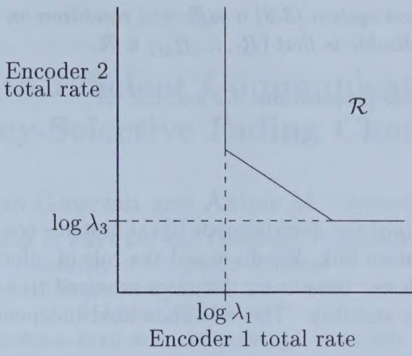


Figure 4: Rate region for example 6.1

Proof: Follows directly from proposition 2.1. □

We now show that for any rate vector in $\mathcal{R}$ there exists an encoder such that the system is uniform in control asymptotically observable.

Proposition 6.2 *Given system (2) a sufficient condition on the rate so that the system is uniform in control asymptotically observable is that* $(\underline{R}_1, ..., \underline{R}_M) \in \mathcal{R}$.

Proof: We prove it for the case when $M = 2$. The proof for the more general case is straightforward. In this case there are three sets of interest: $\Lambda_1 \cap \Lambda_2$, $\Lambda_1 - (\Lambda_1 \cap \Lambda_2)$, and $\Lambda_2 - (\Lambda_1 \cap \Lambda_2)$. By proposition 4.1 we know that the decoder can asymptotically observe the subspace associated with $\Lambda_1 - (\Lambda \cap \Lambda_2)$ with $\sum_{\lambda \in \Lambda_1 - (\Lambda_1 \cap \Lambda_2)} \max\{0, |\lambda|\}$ bits transmitted only by encoder 1. Similarly the decoder can asymptotically observe the subspace associated with $\Lambda_2 - (\Lambda_1 \cap \Lambda_2)$ with $\sum_{\lambda \in \Lambda_2 - (\Lambda_1 \cap \Lambda_2)} \max\{0, |\lambda|\}$ bits transmitted only by encoder two.

This leaves us with $\Lambda_1 \cap \Lambda_2$. Clearly we don't need both encoder one and encoder two sending the same information. Thus any splitting of the rate between encoder 1 and encoder 2 needed to describe $\Lambda_1 \cap \Lambda_2$ is sufficient for the decoder to asymptotically observe the subspace associated with $\Lambda_1 \cap \Lambda_2$. We only require that the combined rate used to describe $\Lambda_1 \cap \Lambda_2$ be greater than $\sum_{\lambda \in \Lambda_1 \cap \Lambda_2} \max\{0, \log |\lambda|\}$. For example encoder one can send coarse bits while encoder two sends fine bits (i.e. most significant digits and least significant digits.) □

Note the relationship to Slepian-Wolf coding. The following example computes a typical rate-region.

Example 6.1 *Let* $A = diag[\lambda_1, \lambda_2, \lambda_3]$ *where* $\lambda_i > 1 \quad i = 1, 2, 3$. *Let* $C_1 = [1, 1, 0]$ *and* $C_2 = [0, 1, 1]$. *The rate region* $\mathcal{R}$ *is shown in figure 4.*

Now we provide lower and upper bounds for asymptotic stability.

Proposition 6.3 *Given system (3.3) a necessary condition on the rate so that the system is asymptotically stabilizable is that* $(\underline{R}_1, ..., \underline{R}_M) \in \mathcal{R}$.

Proof: Follows from proposition 2.2. □

Proposition 6.4 *Given system (3.3) a sufficient condition on the rate so that the system is asymptotically stabilizable is that* $(\underline{R}_1, ..., \underline{R}_M) \in \mathcal{R}$.

Proof: Follows from the propositions 4.2 and 6.2. □

7 Summary

In this paper we examined the deterministic linear systems control problem with a noiseless digital communication link. We discussed the role of information patterns.

We first provided lower bounds on the rates required to achieve asymptotic observability and asymptotic stability. These bounds hold independently of the information pattern chosen.

To compute upper bounds we explicitly described the encoder, decoder, and controller schemes. We characterized two different encoder structures based on whether the encoder observed the control signals or not. Under the use of a primitive quantizer we showed that encoders in encoder class one can achieve these lower bounds. For encoders in encoder class two weaker bounds were provided. These schemes work by keeping track of the uncertainty set of both the error and the state.

We then extended these results to systems with multiple sensors. We showed that the problem reduces to a Slepian-Wolf coding problem.

In the full paper we examine the general problem of performance. We give structural results on the optimal controller for the LQ problem under a rate constraint.

References

[Son] E. Sontag, *Mathematical Control Theory, 2nd Edition.* New York. Springer-Verlag, 1998.

[Tat] S. Tatikonda, *Control under Communication Constraints.* M.I.T. Ph.D. thesis, September 2000.

[Wit] H. Witsenhausen, "Separation of Estimation and Control for Discrete Time Systems." Proceedings of the IEEE, Volume 59, No. 11, November 1971.

On Bandwidth-Efficient Communication Over Frequency-Selective Fading Channels

Ashwin Ganesan and Akbar M. Sayeed
Department of Electrical and Computer Engineering
University of Wisconsin–Madison
1415 Engineering Drive, Rm. 3617
Madison, WI 53706
`ganesan@cae.wisc.edu`, `akbar@engr.wisc.edu`

Abstract

We present a general framework for the design and analysis of transceiver structures for single-transmitter single-receiver multipath fading channels. We assume that the receiver has perfect CSI (Channel State Information), while the transmitter has knowledge of only the channel statistics. Our analysis is based on a virtual MIMO channel characterization in terms of the signal space dimensions. Specific signaling and reception schemes are represented as special cases of the generic MIMO model, and the performance of the schemes is evaluated. The main performance measure is outage capacity. We also present a simple transceiver design whose performance provides a lower bound on the outage capacity of the full-dimensional multipath channel.

1 Introduction

We consider spread-spectrum signaling over a frequency-selective, slow fading channel with impulse response $h(\tau)$. The baseband received signal $y(t)$ can be generally expressed as $y(t) = \int_0^{T_m} h(\tau)x(t-\tau)d\tau + w(t)$, where $x(t)$ is the transmitted signal of duration T, two-sided bandwidth B, $w(t)$ is AWGN (additive white Gaussian noise) with spectral density $\sigma^2 = 1$, and T_m denotes the multipath spread of the channel. There are $L = \lceil T_m B \rceil$ resolvable multipath components. We consider an uncorrelated scattering, Rayleigh fading channel, where the delay coefficients $\{h_l\}_{l=0}^{L-1}$ are uncorrelated, complex zero-mean Gaussian random variables. Average channel power is normalized to unity.

Since the dimensionality of the signal space is $N \approx TB$, an equivalent matrix formulation of the system can be obtained by projecting onto N basis waveforms that capture the sufficient statistics. The system can be viewed as an N-input N-output system in the signal space and represented in the form $\mathbf{y} = \mathbf{Hx} + \mathbf{w}$, where $\mathbf{x}$ is the transmitted signal vector, $\mathbf{y}$ is the received signal vector, $\mathbf{w}$ is AWGN, and $\mathbf{H}$ is the channel matrix represented with respect to the basis. The choice of basis is important from the viewpoint of developing insights into the channel structure and facilitating the performance analysis of transceiver designs.

2 Virtual MIMO Channel Characterizations

By projecting onto a different set of basis waveforms, we obtain four different channel characterizations - the *circulant*, *parallel*, *time-domain* and *frequency-domain* channel models. Let the transmitted signal be represented in the circulant model by the vector $\mathbf{x}^c$, in the parallel model by $\mathbf{x}^p$, in the time domain by $\mathbf{x}^t$, and in the frequency domain by $\mathbf{x}^f$. Similarly for the received signal $y(t)$ and AWGN $w(t)$. We first obtain relationships between the four domains. $\mathbf{x}^t$ and $\mathbf{x}^f$ are simply the samples of $x(t)$ in the time and frequency domains, respectively. They are thus related by $\mathbf{x}^f = \mathbf{U}^H\mathbf{x}^t$, where $\mathbf{U}^H$ is the normalized DFT (Discrete Fourier Transform) matrix, $\mathbf{U}^H_{ij} = e^{-2\pi ij/N}/\sqrt{N}$ for i,j=0,1,...,N-1. Another appealing choice of basis functions for the N dimensional space is a set of spreading waveforms. In a DS-CDMA (Direct Sequence Code Division Multiple Access) system, the symbol is modulated onto a spreading waveform $q(t)$ and transmitted. Here, $q(t) = \sum_{n=0}^{N-1} c_n v(t - n/B)$, where $\{c_n\} \in \{+1, -1\}$ is the spreading code, $v(t)$ is the chip waveform of duration $1/B$ and N is the processing gain. For an arbitrary spreading waveform $q_0(t)$ with code $\mathbf{q}_0^T = [c_0, c_1, \ldots, c_{N-1}]$, define $q_n(t)$ for $n = 1, \ldots, N-1$ to be the nth circularly time-shifted version of $q_0(t)$, i.e., $q_n(t)$ has the spreading code $\mathbf{q}_n^T = [c_{N-n}, c_{N-n+1}, \ldots, c_{N-n-1}]$. Construct the code matrix $\mathbf{Q} = [\mathbf{q}_0, \mathbf{q}_1, \ldots, \mathbf{q}_{N-1}]$. Typically $\mathbf{Q}$ is full-rank, i.e. the basis $\{q_n(t)\}_{n=1}^{N-1}$ is linearly independent and forms a complete basis for the space. The transmitted signal $x(t)$ can be represented in the circulant domain by $\mathbf{x}^c$, where $x(t) = \sum_{n=0}^{N-1} x_n^c q_n(t)$. Sampling at rate $1/B$, we obtain the relation $\mathbf{x}^t = \mathbf{Q}\mathbf{x}^c$. $\mathbf{Q}$ is a circulant matrix with eigendecomposition $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}^H$ for some diagonal $\mathbf{D}$. We define the signal in the parallel domain by the relationship $\mathbf{x}^p = \mathbf{U}^H\mathbf{x}^c$.

The discrete-time representation of the frequency-selective channel is given by: $\mathbf{y}^t = \mathbf{H}\mathbf{x}^t + \mathbf{w}^t$, where $\mathbf{H}$ is a toeplitz matrix. For $N \gg L$, we make the approximation that $\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H$ is circulant. Here, $\boldsymbol{\Lambda} = \mathtt{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{N-1})$, where $\lambda_k = \sum_{n=0}^{L-1} h_l e^{-j2\pi kl/N} = H(f)|_{f=2\pi k/N}$ are samples of $H(f)$ and are zero-mean complex Gaussian with variance 1. It is easy to verify the following N input-N output relationships with respect to the four sets of basis functions: $\mathbf{y}^t = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H\mathbf{x}^t + \mathbf{w}^t$, $\mathbf{y}^c = \mathbf{U}(\mathbf{D}^H\boldsymbol{\Lambda}\mathbf{D})\mathbf{U}^H\mathbf{x}^c + \mathbf{w}^c$, $\mathbf{y}^p = (\mathbf{D}^H\boldsymbol{\Lambda}\mathbf{D})\mathbf{x}^p + \mathbf{w}^p$, and $\mathbf{y}^f = \boldsymbol{\Lambda}\mathbf{x}^f + \mathbf{w}^f$.

3 Transceiver Structures

By using the appropriate MIMO channel model, expressions for ergodic and outage capacity of the multipath channel and of specific transceiver designs can be obtained explicitly in terms of key system parameters – the number of parallel channels created by the transceiver structure, number of occupied signal space dimensions, and the level of diversity exhibited by the channel. While the ergodic capacity of the multipath channel is independent of the number of diversity paths L, the outage capacity improves with increasing L. In particular, the outage capacity can be approximated by a simple transceiver structure: transmit $\mathbf{x}^c = [x_0, 0, \ldots, x_L, 0, \ldots, x_{(N/L-1)L}, 0, \ldots, 0]$ with overall power P, and receive in all N dimensions. The signal occupies N dimensions; there are $\lfloor N/L \rfloor$ orthogonal, parallel channels, each corresponding to a Lth order diversity system. For this scheme, the outage or failure probability for rate R (in bits) is given by [1]

$$P(\mathtt{capacity} < R) = \frac{\gamma(L, \frac{N}{P}(2^{\frac{RL}{N}} - 1))}{\Gamma(L)}. \tag{1}$$

where $\gamma(a, x)$ is the incomplete gamma function $\int_0^x u^{a-1}e^{-u}du$. This expression provides a lower bound on the performance of the full-dimensional multipath channel, and the bound becomes tighter as the dimensionality of the signaling space increases. For constrained transmit power, increasing the number of parallel channels beyond a certain number provides diminishing returns in capacity, while increasing the level of diversity beyond a certain number provides diminishing returns in the probability of error. Our virtual MIMO framework jointly addresses these issues and facilitates evaluating the tradeoffs between various transceiver structures.

References

[1] A. Ganesan and A. M. Sayeed, "A Virtual MIMO Framework for Multipath Fading Channels," *to be submitted for publication*, 2000. `http://dune.ece.wisc.edu/`

Iterative QR Detection for an Uncoded Space-Time Communication Architecture

Mohamed Oussama Damen

Univ. of Minnesota, ECE Dept.
4-174, 200 Union St. S.E. 55455 Minneapolis, MN
damen@ece.umn.edu

Karim Abed-Meraim and Sergiy Burykh

Dept. of Signal Processing
Ecole Nationale Supérieure des Télécommunications de Paris
46 rue Barrault, 75013 Paris.
{abed, burykh}@tsi.enst.fr

Abstract

We study an iterative detection algorithm of an uncoded multi-transmitter multi-receiver system. The main data stream is demultiplexed into M substreams, and each substream is modulated independently then transmitted by its dedicated antenna. The receiver disposes of $N \geq M$ antennas. Over each receive antenna the signal is a superposition of the M substreams affected by independent fades and disturbed by AWGN. The detection algorithm is based on the QR decomposition of the channel transfer matrix which is then used to perform hard or soft inter-substream interference cancellation. Comparisons are done with the V-BLAST optimal order (OPT) detection algorithm. The proposed algorithm is 4 to 8 times less complex than the V-BLAST OPT, and maintains comparable performance.

1 Introduction

Recently, the huge increase in capacity when using multi-antenna compared to single antenna systems in a Rayleigh fading environment [1] has motivated a lot of works. In a system of M transmitters and N receivers, the receive diversity equals to N and it is contained in the uncoded system, while the transmit diversity equals to M and in order to exploit it, one needs space-time codes [2]. In the sequel we only consider the uncoded multi-antenna system. In a pioneering work [3], Bell labs. provided a multi-antenna prototype that reaches very high spectral efficiencies by the use of the V-BLAST (Vertical Bell Labs Layered Space-Time) detection algorithm which realizes a good tradeoff between complexity and performance. In a previous work [5] we have shown that the lattice representation of the multi-antenna environment and the use of the sphere decoder algorithm fully exploit the receive diversity.

In this paper we focus on the complexity issue and propose an iterative detection algorithm based on two QR decompositions of the channel transfer matrix which are then used iteratively to perform hard or soft inter-substream interference cancellation.

The paper is organized as follows: The system model and a brief review of the V-BLAST algorithm are given in Section 2. In Section 3, we present the QR iterative algorithm with hard and soft decisions. Section 4 gives general remarks on the proposed algorithm. Simulation results are given in Section 5. In Section 6 we conclude the paper and give some axis for further research.

2 System model

Consider the system of M transmit and N receive antennas. The single data stream in the input is demultiplexed into M substreams, and each substream is modulated independently then transmitted by its dedicated antenna. It is assumed that the same constellation is used for all the substreams. The transmission is done by burst of length l over a quasi-static Rayleigh fading channel changing randomly every l symbol durations. The power launched by each transmitter is proportional to $1/M$ so that the total radiated power is constant and independent of M. The received signal is given by

$$\mathbf{r} = \mathbf{H}\mathbf{a} + \mathbf{n}, \tag{1}$$

where $\mathbf{a} = (a_1, \ldots, a_M)^T$ denotes the transmitted vector which belongs here to the constellation QAM, and $\mathbf{n}$ is an $N \times 1$ complex AWGN vector, component-wise independent with a variance σ^2 per real dimension. $\mathbf{H}$ is an $N \times M$ transfer matrix of the channel with entries h_{kj}, where h_{kj} is the fading between transmitter j and receiver k modeled by a complex random Gaussian variable of variance 0.5 per real dimension.

Brief summary of the V-BLAST OPT detection algorithm:
The detection is performed as follows [3]:

- Initialization: $i \leftarrow 1$. $\mathbf{G}_1 = \mathbf{H}^+$.
- Recursion:

$$\begin{aligned}
k_i &= \arg(\min_{j \notin \{k_1,\ldots,k_{i-1}\}} \|(\mathbf{G}_i)_j\|^2) \\
\mathbf{w}_{k_i} &= (\mathbf{G}_i)_{k_i} \\
y_{k_i} &= \mathbf{w}_{k_i}\mathbf{r}_i \\
\hat{a}_{k_i} &= Q(y_{k_i}) \\
\mathbf{r}_{i+1} &= \mathbf{r}_i - \hat{a}_{k_i}(\mathbf{H})_{k_i} \\
\mathbf{G}_{i+1} &= \mathbf{H}^+_{\overline{k_i}} \\
i &\leftarrow i+1,
\end{aligned} \tag{2}$$

  where $(\mathbf{G}_i)_{k_i}$ is the k_i-th row of $\mathbf{G}_i$, and where the lowerscript $\overline{k_i}$ refers to zeroing columns $k_1, \ldots, k_i$ because the corresponding components are already canceled. The upperscript $^+$ refers to the pseudoinverse, $Q(\cdot)$ refers to the decision taken on the considered symbol [3].

In a rich scattering environment, V-BLAST algorithm succeeds in separating the desired signal from the multitude of interferers. Nevertheless, we would like to remark the following: At different steps the V-BLAST computes pseudoinverses of matrices of rank values going from M to 1. When the columns of the matrix $\mathbf{H}$ are nearly dependent its pseudoinverse is quite sensitive meaning that small changes in $\mathbf{H}$ can induce arbitrarily large changes in $\mathbf{H}^+$. As for rank deficient cases ($rank(\mathbf{H}) < M$), the sensitivity of the pseudoinverse grows linearly with $M - rank(\mathbf{H})$ (see [6], p. 243). Also, computing M pseudoinverses implies that the complexity of the original V-BLAST is $\approx O(M^4)$ complex arithmetical operations.

Improvement over this algorithm was done in [4], where the nulling process was performed by the use of unitary transformations. This approach reduces the complexity to $29/3M^3$ for $M = N$, and increases the numerical stability of the V-BLAST algorithm.

In the sequel we develop an algorithm of lower complexity, comparable performance and numerical stability when compared to the improved version of the V-BLAST.

3 QR iterative detection algorithm

Let us consider the equation (1). The QR decomposition of the matrix $\mathbf{H}$ is defined as follows

$$\mathbf{H} \;=\; (\mathbf{Q}_1 \; \mathbf{Q}_1') \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{pmatrix}, \tag{3}$$

where $\mathbf{Q}_1$, $\mathbf{Q}_1'$ are $N \times M$ and $N \times N - M$ unitary matrices, and $\mathbf{0}$ is an $N - M \times M$ matrix with all entries equal to zeros, and $\mathbf{R}_1$ is an $M \times M$ upper triangular matrix[1]. In the same manner one obtains

$$\mathbf{H} = (\mathbf{Q}_2' \; \mathbf{Q}_2) \begin{pmatrix} \mathbf{0} \\ \mathbf{R}_2 \end{pmatrix}, \tag{4}$$

with $\mathbf{Q}_2$ an $N \times M$ unitary matrix and $\mathbf{R}_2$ an $M \times M$ lower triangular matrix. Then from (1) it follows

$$\mathbf{y}_1 \;=\; \mathbf{Q}_1^H \mathbf{r} = \mathbf{R}_1 \mathbf{a} + \mathbf{n}_1, \tag{5}$$

$$\mathbf{y}_2 \;=\; \mathbf{Q}_2^H \mathbf{r} = \mathbf{R}_2 \mathbf{a} + \mathbf{n}_2, \tag{6}$$

with $\mathbf{n}_i$ an AWGN of variance σ^2 per real dimension for $i = 1, 2$. The detection is performed as follows: We start by resolving the upper triangular linear system in (5). The symbols are detected successively from a_M till a_1, and the successive interference cancellation at the j-th symbol is performed as follows

$$z_j \;=\; y_{1j} - \sum_{k=j+1}^{M} r_{1,jk} \hat{a}_k = r_{1,jj} a_j + \tilde{n}_{1,j}, \quad j = M \ldots 1, \tag{7}$$

where $\hat{a}_k$ is a soft or hard decision on the symbol a_k for $k = M \ldots j+1$, and $\tilde{n}_{j,1}$ is a non Gaussian noise resulted from the AWGN $n_{1,j}$ and the erroneous decisions on the symbols $a_M \ldots a_{j+1}$. Along with the SIC process we store soft values of the detected symbols: (L_r^j, L_i^j) for the real and imaginary parts of the symbol a_j, $j = M \ldots 1$ which are computed supposing that the noise $\tilde{n}_{j,1}$ is Gaussian. In the next iteration we use equation (6), and the symbols are detected successively from a_1 to a_M. At the step j we compute a new soft value of the detected symbol a_j and compare it to the stored one, then we can either average the two obtained soft values or keep the most reliable one that has the highest absolute value.

---

[1] $\mathbf{R}_1$ is the Cholesky factorization of $\mathbf{H}^H\mathbf{H}$ with $\mathbf{H}^H$ the conjugate transpose of $\mathbf{H}$.

Computing the decision $\hat{a}_j$: Consider equation (7), the soft decision on the symbol a_j is computed as the *a posteriori* statistical expectation and is given by

$$\hat{a}_j \;=\; E[x|z_j] = \sum_{x\in\text{QAM}} \Pr\{x|z_j\}x, \tag{8}$$

which for the 4-QAM constellation has the simple expression

$$\hat{a}_j \;=\; \tanh\left(L_r^j\right) + \sqrt{-1}\;\tanh\left(L_i^j\right), \tag{9}$$

where

$$L_r^j \;=\; \Re\left(\frac{2}{\sigma^2} r_{1,jj}^* z_j\right), \tag{10}$$

and

$$L_i^j \;=\; \Im\left(\frac{2}{\sigma^2} r_{1,jj}^* z_j\right), \tag{11}$$

where $\Re$, $\Im$ denote the real and imaginary parts respectively, and the upperscript * denotes the conjugate. Hard decision on the symbol a_j in the case of 4-QAM is given by

$$\hat{a}_j \;=\; \operatorname{sign}\left(L_r^j\right) + \sqrt{-1}\;\operatorname{sign}\left(L_i^j\right). \tag{12}$$

In simulations with 4-QAM, we update (L_r^j, L_i^j) by choosing the one with the highest absolute value.

Ordering the detected statistics: Similarly to the V-BLAST OPT detector and the SIC type multi-user detectors [7], the best way to perform the SIC process is in the order of decreasing SNR corresponding to the interferers contribution to the received signal. Hence we start by subtracting the most powerful one and so on. In equations (5) and (6), the SNR of the contribution of the symbol a_j to the received signal equals to

$$||\mathbf{H}_j||^2 \;=\; ||\mathbf{R}_{1,j}||^2\frac{\bar{E}_s}{2\sigma^2} = ||\mathbf{R}_{2,j}||^2\frac{\bar{E}_s}{2\sigma^2}, \tag{13}$$

where $||\mathbf{H}_j||$ is the norm of the j-th column of $\mathbf{H}$, $||\mathbf{R}_{i,j}||$ is the norm of the j-th column of the matrix $\mathbf{R}_i$, $i = 1, 2$. $\bar{E}_s$ is the average energy per symbol of the considered QAM constellation.

Since in the proposed algorithm one performs a backward detection (using $\mathbf{y}_1$, $\mathbf{R}_1$) and a forward detection (using $\mathbf{y}_2$, $\mathbf{R}_2$), ordering the columns of $\mathbf{R}_1$, $\mathbf{R}_2$ in an increasing or decreasing way with respect to their square norms improves the performance. On the other hand, it is in the second iteration (forward) that the soft values change and eventually an erroneously detected symbol in the first iteration could be corrected. So it is preferable to make the forward detection optimal by ordering the columns of $\mathbf{R}_1$, $\mathbf{R}_2$ in a decreasing way. This order is verified by our simulations to be optimal (see Section 5). Ordering the columns of $\mathbf{R}_1$, $\mathbf{R}_2$ is done by permuting first the columns of $\mathbf{H}$ then performing up and down QR decompositions on the resulted matrix.

The previous algorithm is similar to the Cholesky-iterative detector described in [8] for multi-user detection. We note the following differences:

1. The optimal ordering of the columns of $\mathbf{R}_1$, $\mathbf{R}_2$.
2. In the case of multi-antenna communication, there is no need to apply the matched filter followed by the noise whitening filter. The QR decomposition is applied directly on the channel matrix and the received signal.
3. The quasi-static nature of the multi-antenna environment allows an adaptive implementation of the QR algorithm, which is not the case in multi-user communication with long codes.

4 General observations

- Number of iterations and OPT statistics: In an uncoded system most of the information contained in the system is extracted after the second iteration. Further processing yields no significant improvement.

  Ordering the columns of $\mathbf{R}_1$ in an increasing way and those of $\mathbf{R}_2$ in a decreasing way with respect to their square norms implies an optimal backward and an optimal forward detection, but it does not improve the performance of the system through the iterations. This is explained as follows:
  Let $\{l_1, \ldots, l_M\}$ be the order of the columns of $\mathbf{R}_1$, so the backward detection gives successively the symbols $a_{l_M}, \ldots, a_{l_1}$. Since the columns of $\mathbf{R}_2$ are ordered in a decreasing way, with the set $\{l_M, \ldots, l_1\}$, then the forward detection gives successively $a_{l_M}, \ldots, a_{l_1}$. Hence, even though each iteration is optimal, it does not fully take advantage of the previous one because the last detected symbol in an iteration is the one with the highest reliability, and in the next iteration we should subtract its contribution at first.

- Soft values: Note that when the channel is real valued, replacing in (10) and (11) yields the following soft decision of 4-QAM constellation symbol a_j

  $$\hat{a}_j = \tanh\left(\frac{2}{\sigma^2} r_{1,jj} \Re(z_j)\right) + \sqrt{-1}\ \tanh\left(\frac{2}{\sigma^2} r_{1,jj} \Im(z_j)\right), \qquad (14)$$

  which is the combination of soft decisions on two independent BPSK modulations [10].

  For the QAM constellation of size $q > 4$, the statistical expectation can be written as

  $$\hat{a}_j = \sum_{a_r \in S_r, a_i \in S_i} \left(a_r \tanh(a_r L_r^j) + \sqrt{-1}\, a_i \tanh(a_i L_i^j)\right) P_{a_r, a_i}, \qquad (15)$$

  where $0 \leq P_{a_r,a_i} \leq 1$, S_r presents the positive real components and S_i presents the positive imaginary components of the QAM constellation. The main problem remains in updating the soft values through iterations because choosing the one with the highest absolute value does not work any more since the QAM points do not have the same absolute value. We have tried to overcome the latter problem by averaging L_r^j, L_i^j or weighting them through iterations, but without improvement over the first iteration. Finding a good method to update the soft values of (15) for QAM of size $q > 4$ is subject to further investigation.
  Nevertheless, one can still obtain acceptable performance with QAM of size $q > 4$

using the QR detector with one iteration (optimal backward or forward detection) with hard or soft decisions; it is shown that one has a slight loss compared to V-BLAST OPT with an algorithm 8 times less complex in this case (see Section 5).

- Performance analysis: Contrary to the problem of multi-user detection [8], assuming a perfect inter-substream interference cancellation here will give a loose bound on the error probability. In order to show this we analyze the post detection SNR of each substream assuming perfect interference cancellation. First we notice that after the backward and forward detection, and since the random variables h_{jk} are i.i.d. with the same variance 0.5 per real dimension, all the symbols have the same order of diversity. This last observation is also verified by simulations. Combining (5) and (1), the post detection SNR of the first symbol a_1 equals

$$\rho_1 \; = \; r_{1,11}^2 \frac{\bar{E}_s}{2\sigma^2} = ||\mathbf{H}_1||^2 \frac{\bar{E}_s}{2\sigma^2}, \tag{16}$$

which is a chi-square random variable with $2N$ degrees of freedom. This implies that the diversity of the first symbol equals to N [9]. On the basis of the previous remark, one concludes that with the hypothesis of perfect inter-substream interference cancellation, the detected symbols have a diversity order of N, which is the one obtained by the ML detection. Obviously this is not the case: Empirically a diversity $\approx N - M + 1$ is obtained at the output of the QR and V-BLAST OPT detector (see Section 5, and also [3]). A rigorous analysis that takes account of the residual interference as a non Gaussian process is still lacking and remains an open problem.

- Computational cost and adaptive implementation: The computational cost of the algorithm is dominated by that of the QR decompositions. The SIC is linear with M and the computation of the soft values in the case of 4-QAM costs $2M$ multiplications and computations of the function tanh. For other QAM modulations the complexity of computing the *a posteriori* statistical expectation (15) is proportional to the size of the modulation.

  The QR decomposition can be done by applying the Householder unitary transformation directly on $\mathbf{H}$ and $\mathbf{r}$ and it costs $2M^2(N - M/3)$ (see [6], page 212). For $M = N$, the QR iterative detector costs $8/3M^3$ which is approximately 4 times less complex than the improved version of the V-BLAST OPT algorithm [4].

  In a quasi-static fading environment, the channel changes very slowly, what allows for the possibility of adaptive implementation of the QR decomposition while estimating the channel, which may reduce the complexity to $\approx O(MN)$. This topic is under investigation.

- The case where $N < M$: In the case where one has more transmitters than receivers, one can still perform the QR iterative detection. Performing the QR decompositions yields

$$\mathbf{H} \; = \; \mathbf{Q}_1\mathbf{R}_1, \tag{17}$$

with $\mathbf{Q}_1$, an $M \times M$ unitary matrix and $\mathbf{R}_1$ an $M \times N$ upper triangular matrix. $\mathbf{R}_2$ is obtained in a similar manner. The backward (respectively the forward) detection

is done similarly except for the last line in $\mathbf{R}_1$ (respectively the first line in $\mathbf{R}_2$) where one has $N-M+1$ variables, and the corresponding symbols are estimated by an exhaustive search over QAM^{N-M+1} which, however, makes it rapidly prohibitive when $M \gg N$.
Note also that the latter algorithm covers the case of correlated fades in a multi-antenna environment.

- Combination with alternate projections: In order to improve the performance of the QR iterative detector one can combine it with alternate projections or a re-processing stage[2] as follows: First we perform the backward and forward detection steps where the columns of $\mathbf{R}_1$, $\mathbf{R}_2$ are ordered by $\{l_1, \cdots, l_M\}$ to obtain a first estimate $\hat{\mathbf{a}}$ on the information symbols. Then an ML detection, by exhaustive search or the sphere decoder [11], is performed on a set of symbols $a_{l_1}, \cdots, a_{l_j}$, where j is called the order of the alternate projection, while considering the values of the rest of symbols as $\hat{a}_{l_{j+1}}, \cdots, \hat{a}_{l_M}$. One obtains a second estimate $\hat{\hat{a}}_{l_1}, \cdots, \hat{\hat{a}}_{l_j}$ which are inserted in the forward detection to obtain $\hat{\hat{a}}_{l_{j+1}}, \cdots, \hat{\hat{a}}_{l_M}$. The complexity of this method is either exponential or polynomial in j, and as it is shown in [11], a slight improvement in performance is obtained which makes it appealing for small j.

5 Simulation results

We have done simulations for $M = 4$ transmitters, $N = 4$ and $N = 6$ receivers with q-QAM constellation where the average energy per bit is fixed to 1 and so that per symbol $\bar{E}_s = 2(q-1)/3$. The plotted curves present the average error rate as a function of SNR, and the variance σ^2 of the AWGN per real dimension is adjusted by the formula

$$\sigma^2 = \frac{M\bar{E}_s}{2\log_2 q}10^{\frac{-SNR}{10}}, \tag{18}$$

where $\log_2 q$ represents the number of transmitted bits per each constellation point, so that σ^2 represents the AWGN power at bit level. Hence, SNR can be called signal to noise ratio per transmitted bit.

In figure 1 we present the QR iterative detector with hard decisions for $M = N = 4$ with and without OPT for iterations 1 and 2. We notice that a considerable gain is obtained after the second iteration. Observe that in this case, the second iteration with OPT has a slight improvement over the second iteration without ordering what is explained by the fact that the forward and backward detections exploit most of the information in the SIC process. Note that the first iteration with OPT is a bit worse than that without ordering because in the backward OPT we start by subtracting the least powerful user what is worse than the SIC without ordering. Also it is noticed that the performance of the QR second iteration with hard decisions and OPT is a bit worse than that of the V-BLAST OPT detector.

Figure 2 shows the same scheme as the figure 1 but the QR detector uses soft decisions. A gain of 2 to 3 dB is observed over the V-BLAST OPT which is due to the use of soft decisions.

Figure 3 presents the performance of the QR detector with only a forward detection when using 16-QAM constellation, with hard or soft decisions, with or without optimal

---

[2] Like the reprocessing stage associated with the V-BLAST algorithm in [11], Chapter 4.

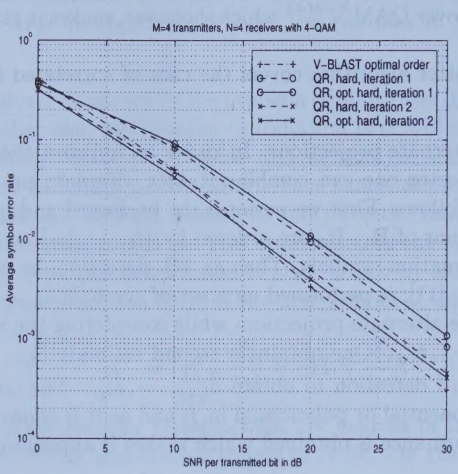


Figure 1: QR detector with hard decisions, M=N=4 with 4-QAM.

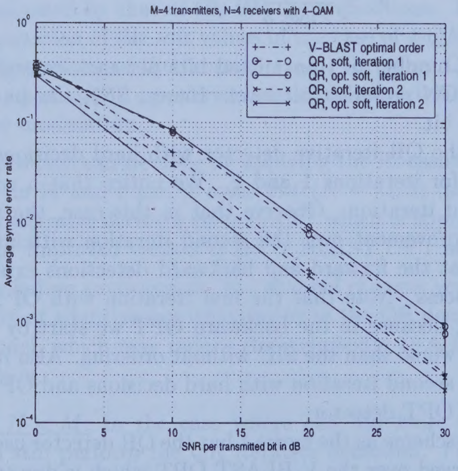


Figure 2: QR detector with soft decisions, M=N=4 with 4-QAM.

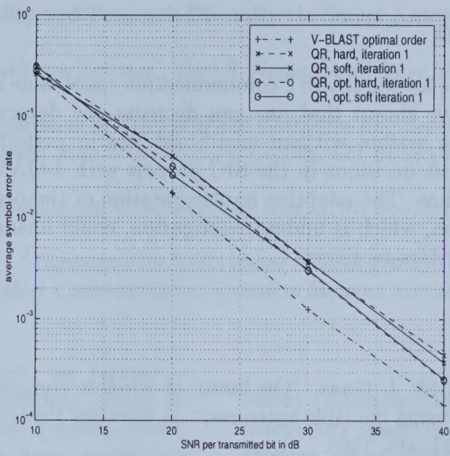


Figure 3: QR detector with hard or soft decisions, M=N=4 with 16-QAM.

order. We notice that soft and hard decisions give the same performance in this case, also we observe $2.5 - 3$ dB of loss compared to V-BLAST OPT, which is a good tradeoff performance-complexity since the QR forward detector is 8 times less complex than V-BLAST OPT.

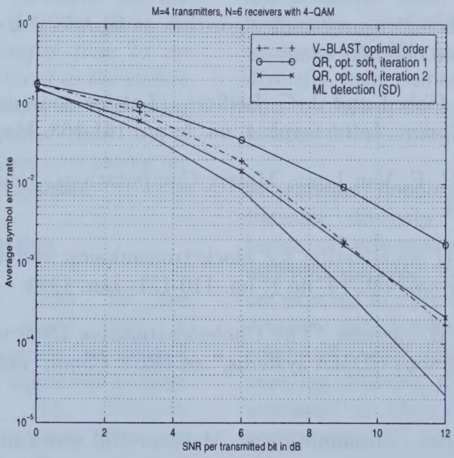


Figure 4: QR detector with soft decisions, M=4, N=6 with 4-QAM.

In figure 4 we present the performance of the QR with soft decisions and OPT of 4 transmitters and 6 receivers with 4-QAM constellation. We also present the performance of the ML detection obtained by the sphere decoder (SD) [5]. It is noticed that, for a fixed number of transmitters, when the number of receivers increases, the performance of the QR detector joins that of the V-BLAST as both approach the performance of the ML detection.

6 Conclusion

A QR iterative detector for the BLAST communication space-time architecture is studied in this paper. We have shown that one can decrease the detection complexity of an uncoded multi-antenna system without significant loss in performance. Also, and as expected, the use of soft decisions in the SIC process with 4-QAM allows for a gain of 2 to 3 dB in performance. The adaptive implementation of the proposed detector along with channel estimation, which is under investigation, could linearize the complexity in the number of transmitters-receivers.

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *in Wireless Personnal Communications*, vol. 6, no. 3, pp. 311-335, 1998.

[2] V. Tarokh, N. Seshadri and A.R. Calderbank, "Space-time codes for high data rate wireless communications: performance criterion and code construction," *IEEE Trans. on Inf. Theory*, vol. 44, no 2, pp. 744-765, March 1998.

[3] G.D. Golden, C.J. Foschini, R.A. Valenzuela and P.W. Wolniasky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *in IEE Elec. Letters*, vol. 35, no. 1, Jan. 1999.

[4] B. Hassibi, "An efficient square-root algorithm or BLAST," Bell labs, Lucent Technologies, http//mars.bell-labs.com.

[5] M. O. Damen, A. Chkeif and J.-C. Belfiore, "Lattice code decoder for space-time codes " *in IEEE Comm. Letters*, vol. 4, no. 5, pp. 161-163, May 2000.

[6] G. H. Golub and C. F. Van Loan, Matrix Computations, John Hopkins University Press, 2nd ed. 1989.

[7] G. Kaleh, "Channel equalization for block transmission systems ," *in IEEE J. on Select. Areas in Comm.* vol. 13, no 1, pp. 110-121, Jan. 1995.

[8] A. Boariu and R. E. Ziemer, "The Cholesky-iterative Detector: A new multiuser detector for synchronous CDMA systems," *in IEEE Comm. Letters*, vol. 4, no. 3, pp. 77-79, March 2000.

[9] J. G. Proakis, Digital Communications, McGraw-Hill series in Electrical and Computer Engineering, 1995.

[10] J. Hagenauer, E. Offer and L. Papke, "Iterative Decoding of Binary Block and Convolutional Codes," *IEEE Trans. on Inf. Theory*, vol. 42, no 2, pp. 429-445, March 1996.

[11] M. O. Damen, "Joint Coding/Decoding in a Multiple Access System, Application to Mobile Communications," Ph.D. Thesis, ENST de Paris, Oct. 1999: http://www.enst.fr/~mdamen

Combined Directionality and Transmit Diversity via Smart Antenna Spatial Sweeping[1]

Seyed Alireza Zekavat[+], *Carl R. Nassar*[+] *and Steve Shattil*[*]

[+]*Department of Electrical and Computer Engineering,*
Colorado State University, Ft. Collins, CO 80523-1373

[*]*Idris Communications, 4980 Meredith Way, #201, Boulder, CO 80303*

[1] *Supported by NASA phase II SBIR grant*
"Development of a Wireless Communication System to Support Airport Surface Operations"

Abstract- *In this paper, we present a novel application for antenna arrays in wireless communication systems. Here, antenna arrays at the base station are used to create (1) directionality which supports spatial multiplexing and (2) a scanning beam pattern, allowing single antenna receivers at the mobile unit to subdivide the symbol time into L partitions, each with independent fades. Combining the signals received in different symbol time partitions (at the mobile unit) results in diversity gains. In this way, diversity benefits (due to an antenna array) are achieved at the mobile unit with the smart antenna array located at the base station (rather than in the mobile itself), a significant cost saving for wireless communication systems.*

I Introduction

Diversity techniques refer to the transmission of a signal over L independent fading channels, ensuring a low probability that all signal components fade simultaneously [1]. Diversity techniques such as frequency, time and space diversity traditionally come at a considerable cost in system bandwidth (frequency diversity), bit rate (time diversity) or expense (spatial diversity) as well as costs in complexity of the receiver structure. The amount of diversity is traditionally chosen based on a compromise between performance and cost.

Novel diversity techniques minimizing complexity have been among the recent antenna array applications. Of particular interest have been *transmit diversity* methods which achieve diversity benefits at the receiving mobile when an antenna array is located at the transmitting base station. For example, in delay diversity [2][3], the transmitter sends N data bearing signals, one per array element, where each element has a unique delay chosen to create an artificially resolvable multipath at the receiver. In [2] and [3], however, the diversity gain comes with a considerable cost to receiver complexity, and, moreover, antenna directionality is no longer available.

In antenna hopping [4], another novel transmit diversity scheme, consecutive bursts of information are transmitted on different antennas. When antenna branches are exposed to different multipath fading, a diversity gain is achieved. Antenna hopping combined with frequency hopping, i.e. applying different frequencies to different antennas, is shown to further enhance performance [5]. In these antenna-hopping schemes, each signal is sent on each antenna in separate portions of time, eliminating the benefit of using of the whole antenna array as a smart antenna. Moreover, in order to achieve path diversity on transmissions from each element, antenna element separation is very high, resulting in large antenna dimensions. In addition, antenna switching produces abrupt periodic phase changes in the received signal, thereby causing periodic errors [6].

Phase sweeping [6] is another form of transmit diversity where a two-branch phase sweeping array at the transmitter induces more rapid fades at the receiver. This improves the channel coding gain in a very slow multipath fading environment. Phase sweeping transmit diversity is an alternative to antenna hoping which avoids the periodic phase changes and corresponding error. However, by using two antennas spatially separated by a large distance it fails to offer the benefits associated with smart antenna schemes (e.g., directionality).

In space-time trellis coding [7], another transmit diversity technique, data is encoded by a channel coder and the encoded data is split into N unique streams, simultaneously transmitted on N transmit antennas. At the receiver, the symbols are decoded using a maximum likelihood decoder. This scheme combines the benefit of forward error correction, coding and diversity transmission to provide considerable performance gains. However, this technique involves additional receiver processing which increases exponentially as a function of bandwidth efficiency (bits/sec/Hz) and the required diversity order; hence, it is not cost-effective for some applications [8]. Moreover, when antenna arrays are used in this fashion, directionality benefits are no longer available.

In this paper, we introduce a novel transmit diversity technique. Here, an antenna array is used to support directionality *and* transmit diversity, via carefully controlled *time varying delays* that result in a slight motion of the beam pattern directed at a mobile user. As a result of the beam motion, time diversity is induced (available) at the mobile receiver *over symbol duration* T_S. This technique represents a novel application for smart antennas. Unlike the methods presented in [2]-[8], this scheme allows a transmit antenna array at the base station to (1) operate as a smart antenna which directs the beam pattern to a mobile user and (2) makes diversity benefits available to the mobile.

In the proposed scheme, increasing L, i.e., diversity gain, is available without cost to bandwidth and bit rate and has only a minimal effect on receiver complexity. This is because the L signals, creating L fold diversity, arrive serially in time (each in time $[iT_S / L, (i+1)T_S / L]$). Hence, a single branch receiver followed by a combining, *which adds terms serially (rather than in parallel),* suffices.

Receivers are designed according to EGC (equal gain combining) and ML (maximum likelihood) criteria. EGC requires no amplitude tracking whereas ML requires such a tracking. The probability of error for both cases is computed analytical and via simulations, and shown to demonstrate large diversity benefits at low costs.

In Section II, the antenna array structure, its delay function and how this delay function ensures independent fades in L partitions of symbol time are introduced. In Section III the receiver structures employing EGC and ML combining are proposed, and in section IV theoretical performance results and simulation results are presented. Section V presents a discussion and conclusions.

II Proposed Antenna Array Structure

Figure 1 presents the delay function applied to each element in the antenna array. As seen in the figure, each adjacent antenna element has delay $\tau(t)$ with respect to its neighboring elements. Now, assuming an antenna array mounted horizontally (on

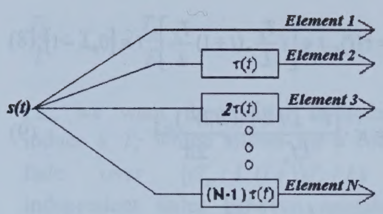


Figure 1: Antenna array structure

the x-axis), the delay elements of Figure 1 generate an antenna pattern in the x-y plane (i.e. $\theta = \pi/2$) corresponding to:

$$Array\ Factor = AF = \frac{\sin\left(\frac{N}{2}\gamma(t)\right)}{\sin\left(\frac{1}{2}\gamma(t)\right)}. \tag{1}$$

Here,

$$\gamma(t) = \frac{2\pi}{\lambda_o} d\cos\phi - \omega_o \tau(t), \tag{2}$$

where λ_o is the wavelength (c/f_o), d is the distance between antenna elements, $\omega_o = 2\pi f_o$ where f_o is the frequency of the transmitted signal, and time t=0 refers to the first instant when the antenna array attempts to contact a mobile user. For simplicity in presentation, it is assumed throughout that the mobile is located at angle $\phi \cong \pi/2$.

With the beam pattern directed toward the desired user, the beam pattern is swept just enough to allow the signals received in L different partitions of symbol duration (T_S) to demonstrate independent fades. This creates full L-fold diversity gains at the receiver. Additionally, the beam pattern is controlled to ensure that the fading *statistics*, such as mean and standard deviation, are constant over the symbol duration. In other words, the base station antenna array sweeps the beam pattern directed to the mobile just enough to create constant large scale fading for the symbol duration, T_S, while ensuring L independent fades within each T_S.

A. Requirement 1: Constant large scale fading statistics

This requirement ensures constant large-scale fading statistics over each symbol time duration T_S. This condition corresponds to

$$\left| T_S \frac{\partial\phi(t,\gamma)}{\partial t} \right| \le (B.W.)_{1/2}, \tag{3}$$

where $(B.W.)_{1/2}$ refers to array half power beamwidth. Equation (3) ensures that the received antenna pattern amplitude is within the 3dB-beamwidth for the entire symbol duration. In other words, if antenna pattern sweeping in T_S is less than $(B.W.)_{1/2}$, we assume that the statistics of large scale fading , i.e., the mean and standard deviation, remain constant in T_S. Now,

$$\left|\frac{\partial\phi(t,\gamma)}{\partial t}\right| = \left|\frac{\partial\phi(t,\gamma)}{\partial\gamma}\cdot\frac{d\gamma(t)}{dt}\right|. \tag{4}$$

and from Equation (2),

$$\left|\frac{\partial\phi(t,\gamma)}{\partial\gamma}\right| = \frac{1}{2\pi d\cdot|\sin\phi|/\lambda_o}. \tag{5}$$

Assuming the usual antenna array design criterion, $d \ge \frac{\lambda_o}{4}$, and assuming the beam pattern directed toward the intended user, i.e., $\phi \cong \pi/2$, (5) results in $\left|\frac{\partial\phi(t,\gamma)}{\partial\gamma}\right| \le 1$. This inequality combined with (4) implies $\left|\frac{\partial\phi(t,\gamma)}{\partial t}\right| \le \left|\frac{d\gamma(t)}{dt}\right|$. Therefore, condition (3) can be replaced by

$$\left| T_S \frac{d\gamma(t)}{dt} \right| = \kappa\cdot(B.W.)_{1/2},\ 0<\kappa\le 1, \tag{6}$$

Substituting (2) into (6) leads to the requirement, $2\pi f_o \cdot T_S \cdot \frac{d\tau(t)}{dt} = \kappa \cdot (B.W.)_{1/2}$, which, solving for $\tau(t)$, reduces to:

$$\tau(t) = \kappa \cdot \frac{R \cdot (B.W.)_{1/2}}{2\pi f_o} \cdot \left(t - \frac{T_S}{2} \right), \qquad (7)$$

where $R = 1/T_S$, and $\frac{T_S}{2}$ is introduced, when solving for $\tau(t)$, to maintain a symmetric movement of the beam pattern around its peak (over T_S). This suggests the selection of delay function $\tau(t)$ which is a linear function of t with slope $\kappa \cdot R \cdot (B.W.)_{1/2} / 2\pi f_o$. In practice, we employ a $\tau(t)$ periodic over symbol time T_S (see below). To maintain a practical $(B.W.)_{1/2}$ value, a small value of κ in the order of 0.01 is used.

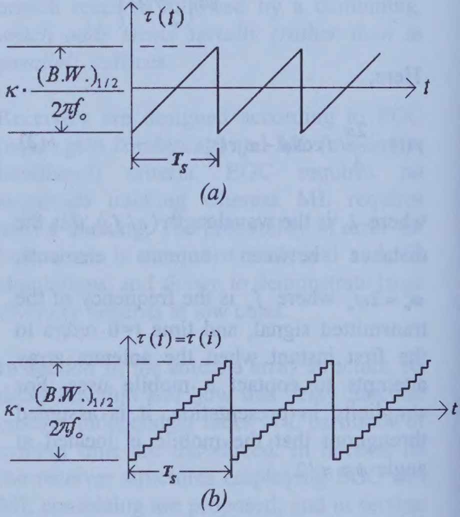
With $\tau(t)$ a linear function of time for each duration, T_S, the final $\tau(t)$ is a saw tooth waveform (Figure 2.a). With this selection of $\tau(t)$, the antenna pattern returns to its initial position and sweeps the same area of space each T_S. Hence, for the i^{th} segment of the symbol time, $[(kT_S + iT_S / L), (kT_S + (i+1)T_S / L)]$, the fades are identical in neigboring symbols (i.e., for symbols, *k, k+1, k+2, …*). This enables fading parameters, e.g. phases, in each segment of T_S to be tracked.

Alternatively we are free to construct the delay function $\tau(t)$ such that it changes discontinuously over T_S. Here $\tau(t)$ is held constant over each T_S / L duration and increases by a fixed step size from one duration to the next (Figure 2.b). In this case, the delay of each element with respect to its neighbors corresponds to:

$$\tau(t) = \tau(i),\ t \in \left[i\frac{T_S}{L}, (i+1)\frac{T_S}{L} \right],\ i \in [0, L-1]; \quad (8)$$

$$\tau(i) = \kappa \cdot \frac{\lceil i - L/2 \rceil}{L f_o} \cdot \frac{(B.W.)_{1/2}}{2\pi}. \qquad (9)$$

The term $\frac{L}{2}$ in $\lceil . \rceil$ is introduced, when solving for $\tau(i)$, to maintain a quasi-symmetric movement of the beam pattern around its peak (over T_S). Here $\lceil x \rceil$ represents rounding x up to the nearest integer.



(a)

(b)

Figure 2. Antenna element delay line a) Continuous and b) Discrete functions

B. Requirement 2: Independent fades

In this section, we determine the requirement which ensures independent fades in each portion $[iT_S / L, (i+1)T_S / L]$. This condition is derived, assuming $\tau(t)$ from equation (7), as follows. First, we want the coherence time, T_o, induced by beam pattern sweeping to correspond to:

$$\frac{T_S}{L} < T_o < 2\frac{T_S}{L}, \tag{10}$$

i.e., we want the spatial sweeping to induce a T_o which allows for a constant fade over $[iT_S / L, (i+1)T_S / L]$ but independent fades between neighboring time intervals of duration T_S / L. The induced coherence time, T_o, corresponds to [9]:

$$T_o \cong \frac{0.423}{f_d}, \tag{11}$$

where $f_d = V_d / \lambda_o$ is the Doppler frequency and V_d, the Doppler speed, refers to the rate of the beam pattern sweeping:

$$\begin{aligned} V_d &= \left|\frac{\partial}{\partial t}(x \cdot \phi(\gamma, t))\right| = \left|x \cdot \frac{\partial \phi(\gamma, t)}{\partial t}\right| \\ &= \left|x \cdot \frac{\partial \phi(\gamma, t)}{\partial \gamma} \cdot \frac{d\gamma(t)}{dt}\right| \\ &= \frac{x \cdot \omega_o}{2\pi d \cdot |\sin \phi| / \lambda_o} \cdot \left|\frac{d\tau(t)}{dt}\right|; \end{aligned} \tag{12}$$

where x is the distance between the base station and the mobile. Again, applying $\phi \cong \pi / 2$ and the $\tau(t)$ in (7) leads to:

$$V_d = \frac{x \cdot \kappa}{2\pi d / \lambda_o} \cdot \frac{(B.W.)_{1/2}}{T_S}. \tag{13}$$

Substituting V_d into f_d, f_d into T_o in (11), and T_o into (10) leads to:

$$1.33\frac{d \cdot L}{x \cdot \kappa} < (B.W.)_{1/2} < 2.66\frac{d \cdot L}{x \cdot \kappa}. \tag{14}$$

Hence, using (14) and knowing the desired degree of diversity (L) (typically L=*20*), the 3dB beamwidth of the smart antenna array is set to ensure the desired diversity gain. Moreover, N, the number of array elements, can be determined using known relationships between $(B.W.)_{1/2}$ and N.

III Receiver Design

A conventional diversity receiver employs parallel branches each collecting one of L signals over duration $[0, T_S]$. In the proposed receiver, diversity is acheieved by combining over $[iT_S / L\ ,\ (i+1)T_S / L]$ portions of symbol time. That is, each of the L diversity components is available by integration with bounds of iT_S / L (lower bound) and $(i+1)T_S / L$ (upper bound). Hence, the proposed diversity receiver requires only one branch, where an output is generated over each $[iT_S / L\ ,\ (i+1)T_S / L]$ interval. The L values generated over duration T_S are stored in a memory box and later combined.

Let the k^{th} information bearing signal entering to the antenna array corresponds to

$$s(t) = \cos(\omega_o t + \theta_k), t \in [kT_S, (k+1)T_S], \tag{15}$$

where $\theta_k = 0 \text{ or } \pi$ contains the information to be conveyed. The transmitted signal out of the antenna, corresponds to

$$s_T(t) = AF \cdot \cos\left(\omega_o t + \theta_k + \frac{N-1}{2}\left(\frac{2\pi}{\lambda_o} d \cos\phi - \omega_o \tau(t)\right)\right), \tag{16}$$

where AF is array factor introduced in (1). The received signal in our proposed system with the mobile at $\phi \cong \pi / 2$ corresponds to

$$\begin{aligned} &r_i(t) = \alpha_i \cos\left(\omega_o\left(t - \frac{N-1}{2} \cdot \tau(t)\right) + \theta_k\right) + n_i(t), \\ &t \in \left[kT_S + i\frac{T_S}{L}, KT_S + (i+1)\frac{T_S}{L}\right], \end{aligned} \tag{17}$$

where $n_i(t)$, $i \in [0, l-1]$ is a white Gaussian noise, α_i is a Rayleigh random variable and perfect phase tracking is assumed.

Additionally it is assumed that beam pattern movement is small enough to ensure $AF \approx 1$. The proposed array structure of section II ensures an independent α_i for different values of i.

Figure 3 presents the optimal receiver for the $r_i(t)$ in (17). In this figure, $v_{L.O.}(t)$ represents the local oscillator term:

$$v_{L.O.}(t) = \sqrt{\frac{T_S}{2}} \cos\left(\omega_o\left(t - \frac{N-1}{2} \cdot \tau(t)\right)\right). \quad (18)$$

Over each interval $[(kT_S + iT_S / L), (kT_S + (i+1)T_S / L)]$ the signal at the output of the integrator in Figure 3 corresponds to

$$r_i = \alpha_i \frac{1}{L}\sqrt{\frac{T_S}{2}} \cos\theta_k + n_i, \quad i = 0,1,\ldots,L-1, \quad (19)$$

where n_i is a zero mean Gaussian random variable with variance σ_n^2 / L. That is,

$$r_i = \alpha_i \frac{1}{L}\sqrt{E_S}\; s + n_i, \qquad i = 0,1,\ldots,L-1, \quad (20)$$

where $s = \cos(\theta_k) \in \{-1,+1\}$ and $\sqrt{E_S} = \sqrt{T_S / 2}$. Following the creation of the L r_i values, the detector can be designed on the basis of equal gain combining (EGC) or maximum likelihood (ML) combining. In EGC detection, the decision variable after combining is:

$$R_{C,EGC} = \sum_{i=0}^{L-1} r_i \,. \quad (21)$$

In ML detection, the decision variable after the combiner corresponds to:

$$R_{C,ML} = \sum_{i=0}^{L-1} \alpha_i r_i \,. \quad (22)$$

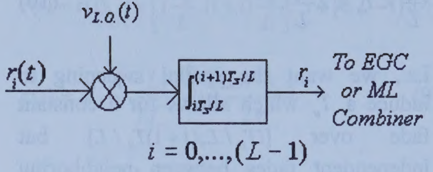


Figure 3: The BPSK induced time diversity receiver

IV Theoretical and simulated performance

A. EGC probability of error

In EGC, the combining of equation (21) is employed, which simplifies to (using r_i in (20)):

$$R_{C,EGC} = \alpha\sqrt{E_S}\; s + n, \quad (23)$$

where:

$$\alpha = \frac{1}{L}\sum_{i=0}^{L-1} \alpha_i \,; \quad (24)$$

$$n = \sum_{i=0}^{L-1} n_i \,. \quad (25)$$

Assuming α_i is a Rayleigh random variable with the probability density function $f(\alpha_i) = \frac{\alpha_i}{\sigma_o^2} e^{-\alpha_i^2 / 2\sigma_o^2}$, then the mean of α corresponds to

$$\mu_\alpha = E(\alpha) = E(\alpha_i) = \sqrt{\frac{\pi}{2}\sigma_o^2}\,, \quad (26)$$

and the variance of α, σ_α^2 is

$$\sigma_\alpha^2 = Var(\alpha) = Var\left(\frac{1}{L^2}\sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\alpha_i\alpha_j\right)$$
$$= \frac{1}{L} \cdot Var(\alpha_i) = \frac{1}{L} \cdot \left(\frac{4-\pi}{2}\right)\sigma_o^2 \,. \quad (27)$$

Additionally for the noise term, n, in (25), the mean remains zero, and the variance corresponds to

$$E(n^2) = L \cdot E(n_i^2) = \sigma_n^2 . \tag{28}$$

Now, assuming we design the proposed system with L sufficiently large (e.g., more than 20), we can apply the central limit theorem: from (24), we can assume α is a Gaussian random variable with the mean and variance characterized by (26) and (27) respectively. Probability of error then corresponds to

$$P(\varepsilon) = \int_{-\infty}^{+\infty} P(\varepsilon|\alpha) f(\alpha) d\alpha , \tag{29}$$

where:

$$P(\varepsilon|\alpha) = Q\left(\frac{\alpha\sqrt{E_S}}{\sigma_n}\right), \tag{30}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-y^2/2} dy , \tag{31}$$

and

$$f(\alpha) = \frac{1}{\sqrt{2\pi\sigma_\alpha^2}} \exp\left(-\frac{(\alpha - \mu_\alpha)^2}{2\sigma_\alpha^2}\right). \tag{32}$$

After substituting (30) and (32) in to (29), and performing mathematical manipulations,

$$P(\varepsilon) = Q\left(\sqrt{\frac{\pi}{4-\pi}} \cdot L \frac{\sqrt{\frac{4-\pi}{2} \cdot \frac{E_S}{N_o} \cdot E(\alpha_i^2)}}{\sqrt{L + \frac{4-\pi}{2} \cdot \frac{E_S}{N_o} \cdot E(\alpha_i^2)}}\right) - Q\left(\sqrt{\frac{\pi}{4-\pi} \cdot L}\right). \tag{33}$$

Here, $N_o = 2\sigma_n^2$ is the noise power.

Letting $E(\alpha_i^2) = 1$, as L increases, the second term in (33) will tend to zero, i.e.,

$$\lim_{L\to\infty} P(\varepsilon) = Q\left(\sqrt{\frac{\pi}{2}} \sqrt{\frac{E_S}{N_o}}\right), \tag{34}$$

which is very close to the $P(\varepsilon)$ for an AWGN channel (Figure 4).

B. Probability of error for ML combining

Using the method presented in [1] for the computation of probability of error in a diversity receiver employing ML combining, and realizing that in our case each branch has signal energy $\sqrt{E_S / L}$ rather than $\sqrt{E_S}$, and each noise component has variance σ_n^2 / L rather than σ_n^2 we determine:

$$P(\varepsilon) = \left[\frac{1}{2}(1-\mu)\right]^L \sum_{l=0}^{L-1} \binom{2L-1}{l} \left[\frac{1}{2}(1+\mu)\right]^l , \tag{35}$$

where

$$\mu = \sqrt{\frac{\gamma_C / L}{1 + \gamma_C / L}}, \ \gamma_C = E\left[\sum_{i=0}^{L-1} \alpha_i^2\right] \cdot \frac{E_S}{N_o} . \tag{36}$$

The theoretical probability of error results for both EGC and ML, with L=25, are provided in Figure 4. Probability of error of ML combining is better than that of EGC, but ML combining requires amplitude tracking which increases receiver complexity.

C. Simulated Performance

Simulation results are performed for an antenna array with spatial sweeping, assuming *L=25* and BPSK transmission. Figure 5 presents performance results. These results confirm how the proposed system, with little change in receiver complexity, moves performance away from poor fading channel performance and toward desired AWGN results. The simulations are in agreement with the theoretical results.

V Discussion and Conclusion

In this paper a novel transmit diversity technique, based on spatial sweeping, has been introduced. By carefully designing the delay elements of base station antenna arrays, we can not only achieve directionality benefit, but, additionally; we induce time diversity at the mobile receiver, significantly increasing receiver performance.

The base station antenna array parameters, namely the relationship between antenna half power beamwidth, L (diversity gain) and the delay (or phase) functions in each antenna element, have been presented. The selections enable time diversity combining in the mobile receiver.

Receivers have been studied employing EGC and ML combining. Theoretical results match simulated performance, and a key feature in receiver design is that time separation between each diversity component allows for a low complexity receiver even in cases of larger degrees of diversity.

References

[1] J.G.Proakis, "Digital Communications," 3^{rd} *Edition, McGraw- Hill,* 1995.

[2] N. Seshadri and J. H. Winters, "Two signaling schemes for improving the error performance of frequency division-duplex transmission system using transmitter antenna diversity," *Int. Journal. Wireless Info. Networks,* vol. 1, no. 1, pp. 49-60, Jan. 1994.

[3] J. H. Winters, "The Diversity Gain of Transmit Diversity in Wireless systems with Rayleigh fading," *in proceeding of the 1994 ICC/SUPERCOMM,* New Orleans, Vol. 2, pp.1121-1125, May 1994.

[4] H. Olofsson, M. Almgren, and M. Hookin., "Transmitter diversity with antenna hopping for wireless communication systems," *Proc. IEEE VTC,* Phoenix, AZ, vol. 3, pp. 1743-1747, May 1997.

[5] Robert W. Heath Jr. and Arogyaswami Paulraj, "Transmit Diversity Using Decision-Directed Antenna Hopping," *in Proc. IEEE Communications Mini-Conference,* pp. 141-145, June, 1999.

[6] A. Hiroike, F. Adachi, and N. Nakajima, "Combined Effects of Phase Sweeping Transmitter Diversity and Channel Coding," *IEEE Trans. On Vehicular Tech.*, Vol. 41, No.2, May 1992.

[7] V. Tarokh, N. Seshadri, and A.R. Calderbank, "Space-time codes for high data rate construction," *IEEE Trans. On Information Theory,* vol. 44, no. 2, pp. 744-765, March 1998.

[8] S. A. Alamouti, V. Tarokh, P. Poon, "Trellis-Coded Modulation and transmit Diversity: Design criteria and performance Evaluation," *in proc. IEEE, 1998 International Conference on Universal Personal Communications.*

[9] J. D. Gibson," Mobile Communication Hand Book," 2^{nd} *Edition, IEEE press, 1999.*

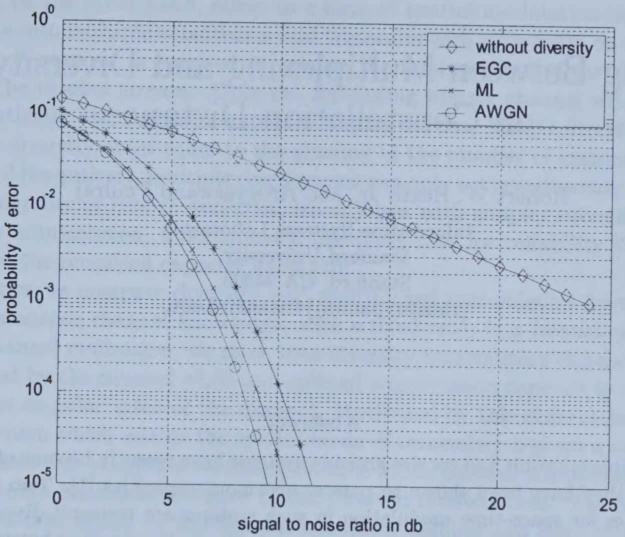


Figure 4. Theoretical results for L=25

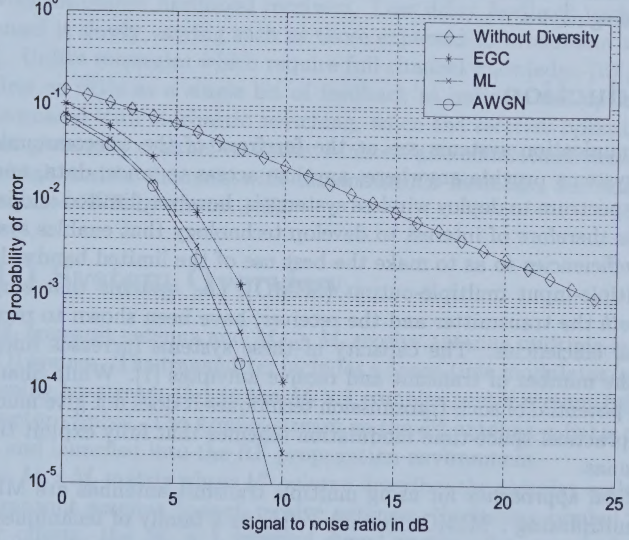


Figure 5: Simulation results for L=25

Switching Between Multiplexing and Diversity Based on Constellation Distance

Robert W. Heath Jr. and Arogyaswami Paulraj *
Information Systems Laboratory
Stanford University
Stanford, CA 94305
{rheath, paulraj}@stanford.edu.

Abstract

Multiple transmit and receive antenna systems have recently become of interest because they have been shown to possess enormous capacities [1]. Two practical techniques for space-time modulation in such systems are transmit diversity and spatial multiplexing. In this paper we propose criteria for choosing between these two schemes given partial channel feedback to the transmitter. In addition we are able to identify channels over which multiplexing is preferred, determined by the channel condition number and the data rate. Simulations demonstrate performance improvements in terms of symbol error rate as a result of switching between the two signaling modes.

1 Introduction

Wireless communication systems are at the forefront of the telecommunications revolution. Such systems provide anywhere, anytime access to voice, data, and multimedia services. The spectrum to deploy wireless systems is, however, limited and is often tightly regulated. It is therefore of interest to develop technology that enables systems to have high spectral efficiencies so as to make the best use of the limited bandwidths available. Recently, multiple-input multiple-output (MIMO), i.e., systems that employ multiple antennas at *both* the transmitter and the receiver, have been shown to promise extraordinary spectral efficiencies. The capacity of these systems increases linearly with the minimum of the number of transmit and receive antennas [1]. While Shannon capacity illustrates the potential of such transmission techniques it does not give much insight into the design of practical space-time modulation schemes that fully exploit the presence of multiple antennas.

Two practical approaches for using multiple transmit antennas are MIMO diversity and spatial multiplexing . MIMO diversity refers to a family of techniques (for e.g., [2], [3], [4], [5], [6]) that attempt to spread information across transmit antennas to enable robust transmission in the presence of fading. These techniques work even with a single receive antenna and thus are also known as transmit diversity techniques. Spatial

---

*Work performed while the authors were at Gigabit Wireless Inc., San Jose, CA. R. W. Heath Jr. is supported by Ericsson Inc. through the Networking Research Center at Stanford University.

multiplexing, on the other hand, refers to a form of spatial modulation that divides the incoming data into multiple substreams and transmits each substream on a different antenna [7] [8]. A receiver is required to remove the mixing effect of the channel and to demultiplex the symbol stream. Over the flat-fading matrix channel with i.i.d. (independent, identically distributed) complex Gaussian entries, MIMO diversity techniques can achieve a diversity order equal to the product of the number of transmit and receive antennas [3]. Alternatively, multiplexing systems can only achieve diversity order at most equal to the number of receive antennas, usually less with simpler suboptimal receivers [9]. Diversity transmission, unfortunately, sacrifices rate for reliability [3] and cannot always achieve the promised capacity gains [10].

In this paper we compare diversity and multiplexed transmission from the point of view of instantaneous channel knowledge, with a fixed rate, in a frequency-flat channel. For a given channel realization, we show that diversity transmission depends only on the power conveyed by the channel while multiplexed transmission depends to a large degree on the smallest singular value of the channel. Motivated by this observation we propose a fixed rate system which adapts the spatial mode of transmission given a low-bandwidth feedback channel. We propose a switching criteria which chooses the modulation scheme with the minimum Euclidean distance of the received constellation. The criteria is also to identify channels, based on the channel condition number and data rate, over which multiplexing will outperform diversity modulation.

Our diversity analysis is applicable to techniques which have equivalent linear representations [11], e.g., Alamouti [2] code, full rate space-time block codes [4]. Our multiplexing analysis is suitable to spatial multiplexing systems without error correction coding and with maximum likelihood receivers. Low delay feedback paths are suitable when the channel is slowly varying such as those expected in broadband wireless access networks [12]. Unlike strategies which require full channel knowledge [13], our adaptive systems requires as little as a single bit of feedback to indicate the optimum mode of operation. Compared with stochastic switching, when the receiver optimizes the transmission scheme given only channel statistics, Monte Carlo simulations demonstrate that, for i.i.d. Rayleigh matrix channels with instantaneous feedback, our adaptive system reduces the required SNR by almost $1dB$.

2 MIMO System Overview

Consider an M_t transmit antenna and $M_r \geq M_t$ receive antenna multiple-input multiple-output (MIMO) system. The transmitter includes a space-time modulator/encoder which maps the incoming bit-stream $\{b_i\}$ to the $M_t \times 1$ complex vector stream $\{\mathbf{s}(n)\}$ at a rate of R bits per vector symbol. The symbol stream on the k^{th} antenna $s_k(n)$ is pulse-shaped, up-converted, and launched into the RF propagation environment.

Let $\mathbf{H}$ be an $M_r \times M_t$ matrix whose k^{th} column describes the complex scalar coefficients from the k^{th} transmit antenna to each receive antenna. Neglecting symbol timing errors and frequency offsets, the $M_r \times 1$ received signal vector after matched filtering and sampling is

$$\mathbf{x}(n) = \sqrt{E_s}\mathbf{H}\,\mathbf{s}(n) + \mathbf{v}(n), \; n = 0, 1, \ldots, N-1 \tag{1}$$

where $\mathbf{v}(n)$ is an $M_r \times 1$ vector realization of a complex circularly symmetric AWGN process with complex distribution $\aleph_c(0, \frac{N_o}{2}\mathbf{I}_{M_r \times M_r})$ and N is the block length. For convenience we normalize the transmitted signal vector to have unit energy, i.e., $\mathrm{tr}(E\{\mathbf{s}(n)\mathbf{s}(n)'\}) =$

1. Block fading, frequency-flat channels, are appropriate when the symbol is much smaller than the channel coherence time and the transmission bandwidth is much less than the coherence bandwidth of the channel.

In this paper we are concerned with two spatial modulation/demodulation methods - MIMO diversity and spatial multiplexing - which are described in the following. Hybrid approaches combining the benefits of both are attractive for large numbers of transmit antennas but have not received much attention and thus will be examined in future work.

MIMO Diversity A general MIMO diversity system is illustrated in Fig. 2. A modulator maps bits to symbols with R bits per symbol. A linear block modulation scheme [14] operates on the N input symbols to produce a codeword $\{\mathbf{s}(n)\}_{n=0}^{N}$. Linear block modulation (described in [11] [15]) incorporates delay diversity [5], linear filtering [5], Alamouti code [2], and full rate space-time block codes [4] as special cases (assuming finite block length). Linear coding schemes can not achieve obtain coding gain [16, p.83] therefore it does not apply to diversity techniques such as space-time trellis coding [3] or layered convolutional codes [17].

Spatial multiplexing A general spatial multiplexing system is illustrated in Fig. 2. A modulator maps bits to symbols with R/M_t bits per symbol. The multiplexer stacks M_t of these symbols to produce the $M_t \times 1$ vector symbol $\mathbf{s}(n)$ for transmission. At the receiver a spatial demultiplexer is applied to $\mathbf{x}(n)$ to extract an estimate of the original bit stream. A number of different receivers algorithms are possible for separation including the maximum likelihood (ML) receiver [9], linear zero-forcing (ZF) or minimum mean-square error (MMSE) receiver [9] and successive cancellation receiver [8]. Due to space limitations in this paper we will only consider the ML receiver. Treatment of the linear receiver can be found in [15].

Observation A key difference between multiplexing and diversity transmission is the symbol modulator. Multiplexing attempts to achieve higher rate by sending independent symbol streams using constellations with fewer points. Diversity, as its name implies, ensures reliable reception by ensuring every symbol on appears every antenna, using an encoder that spreads symbols in space and time. For a given rate R, to determine which scheme is optimal for a given channel we analyze probability of symbol error performance for each scheme.

3 Performance of MIMO Diversity

Let R denote the spectral efficiency of the modulation scheme in bits/s/Hz. Define $M_d = 2^R$ to be number of points in the complex constellation and let $d_{min,d}^2$ be the corresponding minimum distance of the normalized unit energy constellation. Let $d_{min,D}$ be the minimum distance of the diversity constellation at the receiver. It can be shown using the matched filter bound [15] that for linear diversity

$$\begin{aligned} d_{min,D}^2 &\leq \frac{\|\mathbf{H}\|_F^2}{M_t} d_{min,d}^2 &(2)\\ &= \frac{\sum_{k=1}^{M_t}\sum_{m=1}^{M_r}|h_{mk}|^2}{M_t} d_{min,d}^2 &(3)\\ &= \frac{\sum_{k=1}^{M_t}\lambda_k^2}{M_t} d_{min,d}^2 &(4)\\ &\leq \frac{(M_t-1)\lambda_1^2+\lambda_{M_t}^2}{M_t} d_{min,d}^2 &(5)\end{aligned}$$

where we have used the definition of Frobenius norm in (3), a standard relation between the Frobenius norm and the singular values $\{\lambda_k\}$ of $\mathbf{H}$ [18], i.e., $\|\mathbf{H}\|^2 = \sum_k \lambda_k^2$ where λ_k is the k^{th} singular value of $\mathbf{H}$, and the fact that $\lambda_k \leq \lambda_1 \; k = 1, \ldots M_t$. Essentially, $d^2_{min,D}$ gives the effective minimum distance of the constellation at the receiver after processing by the diversity block. Equality in (2) is achieved for the Alamouti scheme [2], full rate space-time block codes [14], and delay diversity with $M_t = 2$ and maximum likelihood sequence estimation [19].

Note that the minimum distance of the received constellation is effectively a scaled version of the per-antenna constellation. Thus, given $\mathbf{H}$, we can upper-bound the exact probability of symbol error for an M_d-point constellation as

$$P_D \;\leq\; N_e \, Q\left(\sqrt{\frac{E_s}{2N_o} d^2_{min,D}}\right) \tag{6}$$

$$< \;(M_d - 1)\, Q\left(\sqrt{\frac{E_s}{2N_o} d^2_{min,D}}\right) \tag{7}$$

where (6) is the nearest neighbor union bound (NNUB) with N_e average nearest neighbors, and (7) is the union bound on the probability of symbol error. When (2) does not hold with equality then (6) and (7) are approximate.

4 Performance of Spatial Multiplexing

Let $\mathbf{x}$ be the received data vector at sample time n, $\mathbf{s}$ the transmitted vector symbol, $\mathcal{S}_{SM}$ the set of all possible transmitted vectors $\mathbf{s}$, and define $M_{sm} = 2^{R/M_t}$. The output of the channel appears in (1). The size of $\mathcal{S}_{SM}$ is $M_{SM} := |\mathcal{S}_{SM}| = M_{sm}^{M_t} = M_d$ which is expected since the spectral efficiency of both signaling schemes is held identical for a fair comparison. The ML estimate of the transmitted vector $\mathbf{s}$ is defined as

$$\hat{\mathbf{s}} \;=\; \arg \min_{s \in \mathcal{S}_{SM}} \left\| \mathbf{x} - \sqrt{\frac{E_s}{M_t}} \mathbf{H}\mathbf{s} \right\|^2 . \tag{8}$$

For small M_{sm} and M_t, say $M_{sm} = 16$ and $M_t = 2$, ML decoding may be of practical interest, but for larger values suboptimal receivers with lower complexity are generally preferred. For spectral efficiencies less than 10 bits/s/Hz the ML receiver may be practically implemented using application specific integrated circuits (ASICs) for moderate data rate applications. For higher spectral efficiencies and/or data rates, other receivers such as the linear receiver or the layered space-time receiver [8] are promising.

Exact error probability of the system can be computed using a geometric approach [20] but the result cannot be readily expressed in closed form. We therefore examine upper bounds on the error probability. The upper bound requires computation of $d^2_{min,SM}$, the squared minimum distance of the received constellation $\{\mathbf{H}\mathbf{s} | \mathbf{s} \in \mathcal{S}_{SM}\}$.

Suppose that $\mathbf{s}_i \in \mathcal{S}_{SM}$, $\mathbf{s}_j \in \mathcal{S}_{SM}$ such that $\mathbf{s}_i \neq \mathbf{s}_j$ belong to the transmit vector constellation. Let the squared minimum distance of the output constellation be defined as

$$d^2_{min,SM} \;:=\; \min_{\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S}_{SM}, \mathbf{s}_i \neq \mathbf{s}_j} \frac{\|\mathbf{H}(\mathbf{s}_i - \mathbf{s}_j)\|^2}{M_t}. \tag{9}$$

Unlike $d^2_{min,D}$ in (2), $d^2_{min,SM}$ not readily a function of $d^2_{min,sm}$, the per-antenna transmit constellation. This is because the symbols $\mathbf{s}_i$ undergo a linear transformation by the channel which does not preserve their distance properties unless $\mathbf{H}$ is unitary. The computation in (9) requires a search over $\mathcal{S}_{SM}(\mathcal{S}_{SM} - 1)$ vectors which is prohibitive for larger constellations. We therefore find it useful to develop both upper and lower bounds on $d^2_{min,SM}$ to simplify computation and to develop intuition on properties of the receive constellation.

First consider the upper bound. To simplify (9) we will compute the minimum squared distance over a smaller set of transmitted symbols $\mathbf{s}_i$. Let $\overline{\mathcal{S}_{SM}}$ be a reduced set of data vectors such that $|\overline{\mathcal{S}_{SM}}| < |\mathcal{S}_{SM}|$. Then it follows that

$$d^2_{min,SM} \;\leq\; \min_{\mathbf{s}_i,\mathbf{s}_j \in \overline{\mathcal{S}_{SM}}} \frac{\|\mathbf{H}(\mathbf{s}_i - \mathbf{s}_j)\|^2}{M_t} \tag{10}$$

which is an upper bound (i.e., optimistic) on the minimum distance since we are searching over a reduced symbol set. A general way of constructing $\overline{\mathcal{S}_{SM}}$ would be to remove vectors from $\mathcal{S}_{SM}$ in such a way that the minimum transmit distance squared is as large as possible. This ensures that the subset of vectors which remains will appear as spatially white as possible. A practical way to construct $\overline{\mathcal{S}_{SM}}$ would be to take points from a smaller per-antenna constellation. For example, with two transmit antennas and 16-QAM per-antenna constellations, $\mathcal{S}_{SM}(\mathcal{S}_{SM} - 1) = 240$. If we search over a space of vectors constructed from 4-QAM then we only need to search over $\overline{\mathcal{S}_{SM}}(\overline{\mathcal{S}_{SM}} - 1) = 12$ vectors, a significant reduction.

Now consider the lower bound. We will compute this bound using some properties from linear algebra. Using the Rayleigh-Ritz theorem [18] that $\|\mathbf{H}(\mathbf{s}_i - \mathbf{s}_j)\|^2 \geq \lambda^2_{M_t}\|(\mathbf{s}_i - \mathbf{s}_j)\|^2$ we have

$$\min_{\mathbf{s}_i,\mathbf{s}_j \in \mathcal{S}_{SM}} \|\mathbf{H}(\mathbf{s}_i - \mathbf{s}_j)\|^2 \;\geq\; \min_{\mathbf{s}_i,\mathbf{s}_j \in \mathcal{S}_{SM}} \lambda^2_{M_t}\|(\mathbf{s}_i - \mathbf{s}_j)\|^2 \tag{11}$$

$$= \lambda^2_{M_t} d^2_{min,sm} \tag{12}$$

where we have defined the minimum squared distance of the per-antenna transmit constellation to be $d^2_{min,sm} = \min_{\mathbf{s}_i,\mathbf{s}_j \in \mathcal{S}_{SM}} \|\mathbf{s}_i - \mathbf{s}_j\|^2$. Clearly, equality occurs if there exists an $\mathbf{s}_i - \mathbf{s}_j$ which is a scalar multiple of the minimum right singular vector of $\mathbf{H}$. Using (9) and (11) we have the desired result that

$$d^2_{min,SM} \;\geq\; \frac{\lambda^2_{M_t}}{M_t} d^2_{min,sm} \tag{13}$$

with equality when there exists a nonzero $\mathbf{s}_i - \mathbf{s}_j$ for which $\mathbf{H}(\mathbf{s}_i - \mathbf{s}_j) = \lambda_{\mathbf{M_t}}$. For $M_r \times M_t$ channels with rank$(\mathbf{H}) < M_t$, (13) indicates that spatial multiplexing may perform quite poorly. This is an important observation because as M_{sm} increases, $\mathcal{S}_{SM}$ becomes more dense and the probability that an error vector lies near $\mathbf{e}_{min}$ increases.

Now consider an upper bound on the probability of symbol error. Let N_i be defined as the number of nearest neighbors to $\mathbf{Hs}_i$ in the receive constellation $\{\mathbf{Hs} | \mathbf{s} \in \mathcal{S}_{SM}\}$ and let $N_e := \frac{1}{|\mathcal{S}_{SM}|}\sum_{k=1}^{\mathcal{S}_{SM}} N_k$ be the average number of nearest neighbors assuming equally likely transmit vectors. Then we obtain the NNUB and the union bound respectively as

$$P_{SM} \;\leq\; N_e Q\left(\sqrt{\frac{E_s}{2N_o} d^2_{min,SM}}\right) \tag{14}$$

$$\leq\; (|\mathcal{S}_{SM}| - 1) Q\left(\sqrt{\frac{E_s}{2N_o} d^2_{min,SM}}\right). \tag{15}$$

In (14), unlike in the diversity case (6), the number of nearest neighbors is also a function of the channel whereas in diversity signaling it is only a function of the constellation. Since computation of N_e has higher complexity than the computation of $d^2_{min,SM}$ in (9) we find that (15) will be more useful in the sequel.

The upper bound developed in (10) does not preserve the direction of the inequality so (15) becomes an approximation. On the other hand, substituting the lower bound in (13) into (15) yields

$$P_{SM} \;\leq\; (|\mathcal{S}_{SM}|-1)Q\left(\sqrt{\frac{E_s}{2N_o}\frac{\lambda^2_{M_t}d^2_{min,sm}}{M_t}}\right). \tag{16}$$

From (16) we see that unlike the diversity case, spatial multiplexing is influenced by the smallest mode in the channel and is therefore sensitive to channel rank. This means that "good" channels are as close to unitary as possible.

5 Adaptive Spatial Modulation

The analysis in the previous sections shows that the performance of each transmission scheme depends on the MIMO channel $\mathbf{H}$. Multiplexing transmission is dependent on the minimum singular value therefore degrades as channel $\mathbf{H}$ becomes ill conditioned. MIMO diversity, however, is only dependent on the channel through the Frobenius norm, therefore diversity works independently of channel condition number. In practical systems, the channel (and therefore its condition number), vary over time. It is thus advantageous to implement an adaptive system, as in Fig. 3, that optimizes the mode of operation depending on the channel realization.

The system in Fig. 3 consists of a transmitter with a switch between a diversity and a multiplexing modulator, a receiver unit with the corresponding pair of receivers, and a low-bandwidth feedback path. When the channel is slowly changing, the receiver computes the optimum signaling technique and sends this information (a single bit) to the transmitter. The transmitter then employs the corresponding spatial signaling method. Low-bandwidth feedback channels, for example, are present in most cellular systems for the purpose of power control. At present we assume perfect channel knowledge at the receiver and zero feedback delay. Delay in the feedback path or channel estimation errors will create decision errors and thus will degrade switching performance.

In this work we focus on determining the optimum signaling strategy for a given rate R. Since optimizing the bit error probability is difficult due to the bit-to-symbol mapping, we instead optimize the symbol error probability probability of symbol error. Clearly, choosing the best mode for a given channel will result in a lower symbol error rate than can be obtained with either diversity or multiplexing alone. This decrease can be used to reduce the required transmit power or improve reliability and is suitable for fixed rate services such as voice and best-effort traffic such as IP (internet protocol). An interesting extension is to optimizing both the rate and the coding scheme for a target symbol error rate. We leave this to future research.

Consider the probability of (vector) symbol error P_D and P_{SM}. Exact computation of P_D is possible for many diversity schemes (e.g., Alamouti [2], space-time block codes [4]) for many constellations. Exact evaluation of P_{SM} is possible but requires multidimensional integration [20]. A solution is to consider the union bound on P_D in (7) and P_{SM} in (15). This is well motivated because, at high SNR, the probability of symbol error is often dominated by the error event with the smallest minimum distance. Since the spectral efficiency of each encoding scheme is the same, we can ignore the multiplicative terms outside the error integral $Q(\cdot)$, and arrive at the following.

Criterion 1 Minimum Euclidean distance based switching *If* $d^2_{min,TD} \le d^2_{min,SM}$ *choose multiplexing, otherwise choose diversity.*

Since the receive constellations are a function of the channel, the value of the minimum distance must be computed for every realization. Minimum distance is computed easily for diversity transmission using (2). For multiplexing transmission with the ML receiver it can be computed from (9). If computation is an issue it is possible to use the lower bound on $d^2_{min,SM}$ in (13). A similar minimum Euclidean distance criterion also has also been applied to select subsets of transmit antennas for spatial multiplexing systems [21].

As Criterion 1 is based on a comparison of upper bounds, it is only approximate. An better comparison would be to compare the NNUB in (6) and (14) which are tight at high SNR. This would capture, at least in part, the additional effect of the weight spectrum of the received constellation. Unfortunately, computation of the NNUB in P_{SM} requires knowledge of the number of nearest neighbors for each point in the received constellation which may be beyond reasonable computation.

6 Channel Characterization

Let us determine when the lower bound on minimum distance for spatial multiplexing in (13) exceeds the minimum distance for diversity transmission in (5). This comparison will classify channels, based on condition number, for which spatial multiplexing will always be better than diversity transmission. Since we have used a lower bound for multiplexing and an upper bound for transmit diversity, there will be other channels not captured by the formula in (18) for which multiplexing is also suitable. Plugging (5) and (13) into the inequality in Criterion 1, and simplifying we have

$$\frac{(M_t-1)\lambda_1^2+\lambda_{M_t}^2}{M_t} d^2_{min,td} \le \lambda_{M_t}^2 d^2_{min,sm}. \qquad (17)$$

Let $K := \lambda_1/\lambda_{M_t}$ be the relative condition number of the channel. Rearranging (17) reveals the following.

Definition 1 Multiplexing preferred channels. *Channels* **H** *with condition number* K *are suitable for multiplexing if*

$$K \le \sqrt{\frac{\frac{d^2_{min,sm}}{d^2_{min,td}}-1}{M_t-1}}. \qquad (18)$$

The significance of (18) is that we can find a condition number threshold which guarantees that multiplexing will outperform diversity transmission.

A immediate result from (18) is that

Lemma 1 *Multiplexing preferred channels do not exist if* $d^2_{min,sm} < d^2_{min,d}$.

Proof: Follows from the fact that regular condition number $K \geq 1$. □ Note that the per-antenna spectral efficiency for multiplexing is R/M while for diversity it is R. Therefore when the same constellation family (e.g., QAM constellations) are employed $d^2_{min,sm} > d^2_{min,d}$ and multiplexing-preferred channels will always exist for $M_t > 1$. With R/M_t fixed and R (and M_t) increasing, the value of this threshold increases provided that $d^2_{min,d}$ decreases more rapidly than M_t grows. This is the case for many practical constellations (e.g., QAM, PSK, PAM) thus multiplexing is preferred for large rates. Further elaboration on this point is available in [15].

7 Simulations

In this section we evaluate the performance the system in Fig. 3 using minimum Euclidean-distance based switching. Let $M_t = 2$, $M_r = 2$ and $R = 4b/s/Hz$. The encoder switches between the Alamouti MIMO diversity scheme [2] with 16-QAM modulation and spatial multiplexing with 4-QAM per-antenna constellation. For comparison with research in diversity literature we use the i.i.d. uncorrelated Rayleigh matrix channel model which implies that $[\mathbf{H}]_{ij} = h_{ij}$ are complex $\aleph_c(0,1)$ such that $Eh_{ij}h^*_{kl} = 0$ for $i \neq k$, $j \neq l$. In the simulation we assume that the channel is constant over a block of $N = 100$ symbols. Perfect channel knowledge and zero-delay feedback is assumed.

In Fig. 4 we plot the probability of vector symbol error, estimated over 2500 Monte Carlo simulations, obtained using minimum Euclidean distance switching and various bounds on $d^2_{min,SM}$. Stochastic switching, that is choosing one mode for all channel realizations, would choose SM up to $SNR = 14dB$ and TD for larger SNRs. We see that minimum Euclidean distance switching follows the SM curve for low SNRs then takes the shape of the TD curve. About $0.75dB$ of switching gain over the best of SM and TD is obtained in higher SNRs. Using the upper bound on $d^2_{min,SM}$ in (10) is a good approximation at lower SNRs but suffers at higher SNRs. The lower bound in (13) gives about $0.25dB$ over TD for higher SNRs and is less exact at lower SNRs. This seems to indicate that there are channels with condition numbers outside the no-diversity region which are also suitable for multiplexing. Results using other switching criteria are reported in [15].

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, March 1998.

[2] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *"IEEE Jour. Select. Areas in Commun."*, pp. 1451 – 1458, Oct. 1998.

[3] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," *IEEE Trans. on Info. Th.*, vol. 44, no. 2, pp. 744–765, March 1998.

[4] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. on Info. Th.*, pp. 1456–1467, July 1999.

[5] A. Wittneben, "Base station modulation diversity for digital simulcast," in *Proc. IEEE VTC*, May 1993, vol. 3, pp. 505–511.

[6] R. W. Heath Jr. and A. Paulraj, "Transmit diversity using decision-directed antenna hopping," in *Proc. of ICC Comm. Theory Mini-Conf.*, June 1999, pp. 141 – 145.

[7] A. Paulraj and T. Kailath, "U. S. #5345599: Increasing capacity in wireless broadcast systems using distributed transmission/directional reception (DTDR)," September 1994.

[8] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multiple antennas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, 1996.

[9] B. A. Bjerke and J. G. Proakis, "Multiple-antenna diversity techniques for transmission over fading channels," in *Wireless Communications and Networking Conference*, Sept. 1999, vol. 3, pp. 1038–1042.

[10] S. Sandhu and A. Paulraj, "Space-time block coding : A capacity perspective," To appear in *IEEE Comm. Letters*.

[11] S. Sandhu, R. W. Heath Jr., and A. Paulraj, "What are optimal space-time block codes?," in *Proc. of Allerton Conf. on Comm. Cont. and Comp.*, October 2000.

[12] K. Sheikh, D. Gesbert, D. Gore, and A. Paulraj, "Smart antennas for broadband wireless access networks," *IEEE Comm. Magazine*, vol. 37, no. 11, pp. 100–105, Nov. 1999.

[13] H. Sampath, P. Stoica, and A. Paulraj, "Generalized precoder and decoder design for mimo transmission systems using the weighted mmse criterion," Submitted to *IEEE Trans. on Communications* April 2000.

[14] G. Ganesan and P. Stoica, "Space-time diversity using orthogonal and amicable orthogonal designs," in *Proc. ICASSP*, 2000.

[15] R. W. Heath Jr. and A. Paulraj, "Fixed rate spatial adaptation in multiple antenna systems," Submitted to *IEEE Trans. Comm.* October 2000.

[16] E. Biglieri, D. Divsalar, P. J. McLane, and M. K. Simon, *Introduction to trellis-coded modulation with applications*, Macmillan Pub. Co., 1991.

[17] R. R. Hammons and K. El Gamal, "On the theory of space-time codes for psk modulaiton," *IEEE Trans. on Info. Theory*, vol. 30, no. 2, pp. 524–542, March 2000.

[18] H. Lutkepohl, *Handbook of Matrices*, John Wiley & Sons, 1996.

[19] J. H. Winters, "The diversity gain of transmit diversity in wireless systems with rayleigh fading," *IEEE Trans. on Veh. Tech.*, vol. 47, no. 1, pp. 119–123, Feb. 1998.

[20] S. Talwar and A. Paulraj, "Blind separation of synchronous co-channel digital signals using an antenna array. ii. performance analysis," *"IEEE Trans. on Sig. Proc."*, vol. 45, no. 3, pp. 706–718, March 1997.

[21] R. W. Heath Jr., S. Sandhu, and A. Paulraj, "Antenna selection for spatial multiplexing systems based on minimum error rate," Submitted to Proc. of Int. Conf. on Comm. 2001, September 2000.

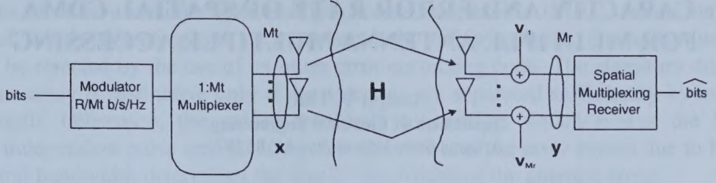

bits
Modulator R/Mt b/s/Hz
1:Mt Multiplexer
Mt
x
H
v1
Mr
vMr
y
Spatial Multiplexing Receiver
bits

Figure 1: A spatial multiplexing system

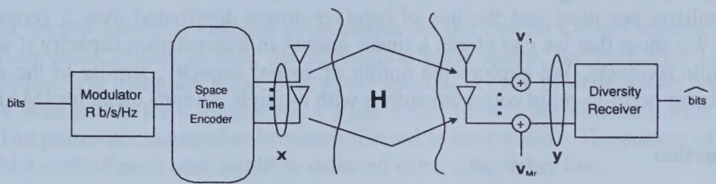

bits
Modulator R b/s/Hz
Space Time Encoder
x
H
v1
vMr
y
Diversity Receiver
bits

Figure 2: A MIMO diversity system

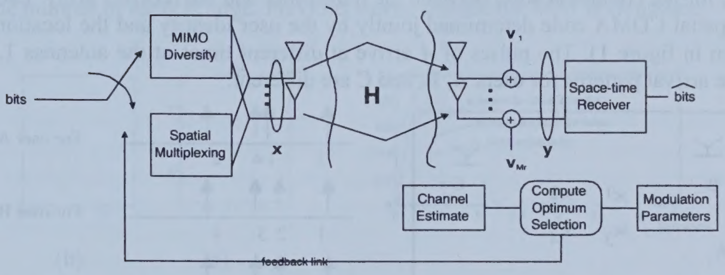

bits
MIMO Diversity
Spatial Multiplexing
x
H
v1
vMr
y
Space-time Receiver
bits
Channel Estimate
Compute Optimum Selection
Modulation Parameters
feedback link

Figure 3: Hybrid spatial modulation system

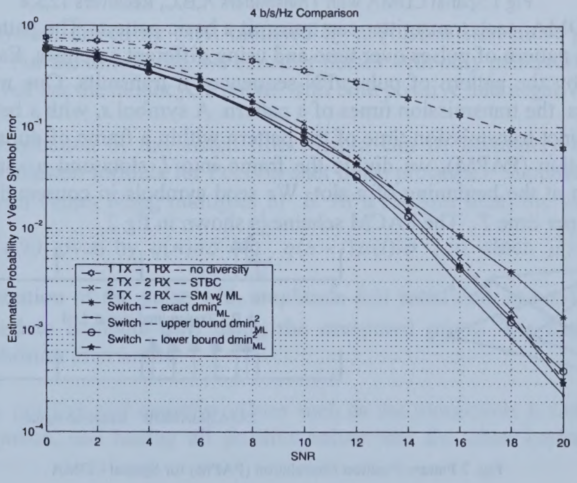


Figure 4: Rate $4b/s/Hz$ comparison

CAPACITY AND ERROR RATE OF SPATIAL CDMA FOR MULTIPLE ANTENNA MULTIPLE ACCESSING

Joseph Y. Hui
Department of Electrical Engineering
Arizona State University, AZ 85287

Abstract

Pulse Radio has been considered a power efficient method for communication in a multiple path fading environment. We extend the method to the use of multiple transmitters per user, and the use of receiver arrays distributed over a geographic area. We show that we can obtain a linear growth in transmission capacity if we use multiple receivers. We explore the notion of spatial capacity, similar to the earlier results for point-to-point communications with multiple antenna elements (MAE).

1. Introduction

We focus on the use of pulse radio [1-4] with MAE. The spatial distribution of MAE, together with the distinct multiple paths between the transmitter and receiver, generates a rich signature for the communication between the transmitter and the receiver array. Each user has a unique spatial CDMA code determined jointly by the user identity and the location of the user (as shown in figure 1). The pulses of *A* arrive at different times at the antennas 1, 2, 3 and 4. The pulse arrival patterns for users A, B, and C are different.

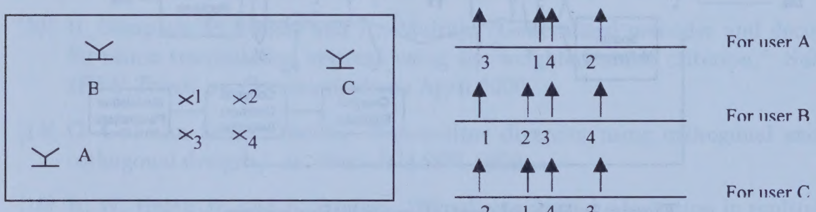


Fig 1.Spatial CDMA with Transmitters A,B,C, Receivers 12,3,4

In spatial CDMA, each transmitter can transmit a basic pattern. The pattern can simply be a single pulse, or a pattern of pulses over time and using different antennas. Each user has its own signature given by the pattern of pulses/PN-sequences it transmits. One method of encoding information is via the transmission times of a pattern. A symbol *s*, with *s* being an integer from 1 to *L*, has a distinct transmission time of the pattern within a frame of duration *T*. For Pattern Position Modulation (PAPM), we divide the frame into *L* slots and a symbol is defined by sending a pattern at the beginning of a slot. We send symbols in consecutive frames. We can send $\log_2 L$ bits per time *T*. The PACM scheme is shown in fig 2.

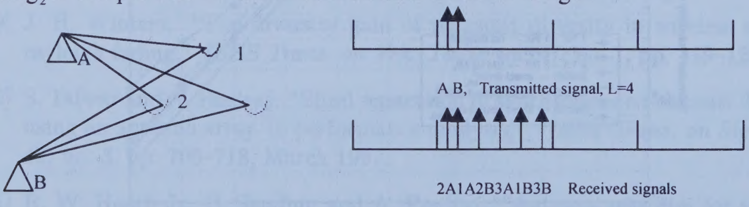


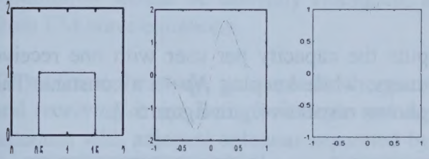
Fig. 2 Pattern Position Modulation (PAPM) for Spatial CDMA

The patterns of other users constitute multiple access interference (MAI), which may result in the event of falsely detecting one or more symbols within a frame. Such spurious false detection may be rejected by the use of an outer error correcting code. The signature due to the receiver antenna array is distinctive only if the antennas are separated sufficiently by multiples of a pulse length. Otherwise, the pulses are likely to overlap, which makes the Poisson assumption of independent pulse arrivals incorrect when we analyze error events due to MAI. In this case, spectral bandwidth determines the spatial bandwidth of the antenna array.

3. Computing the MAI Distribution

We consider several spectrally efficient pulse shapes for our Spatial CDMA schemes shown in figure 3. These pulse shapes include the rectangular pulse shape, the monocycle Gaussian pulse, and the monocycle triangular pulse. The power spectral density of these pulses are shown in figure 4. We assume that for a transmitter, a pulse could be sent per unit time with a small probability *p*. The pulses are assumed to be match filtered at the receiver. The pattern of pulses, the spatial CDMA code of each user, could be detected using a tap delay line.

A key assumption made in our analysis is that the MAI is modeled as a Poisson shot noise process, that the pulse arrivals are independent. This assumption is true only if the antennas are far apart and the number of users is large. Previous works assume the MAI is Gaussian.



(a) (b) (c)

Fig.3 Different pulse shapes, a: rectangular pulse, b: monocycle Gaussian pulse, c: monocyle triangular pulse

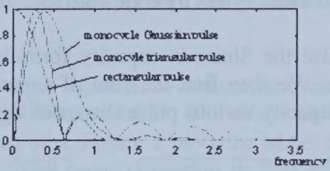


Fig. 4 Spectrum of the three pulse shape

We consider the system driven by MAI modeled by a Poisson shot noise process with a Poisson arrival rate λ. For the time being, we consider pulses with only positive values, such as the rectangular pulse, but not the monocycle pulses with both positive and negative values.

The probability distribution of the match filter output can be accurately and efficiently computed using a convolution formula for Poisson shot noise processes. Suppose all amplitude values are discrete, with values being multiples of δ. At a random time, the sample value of $z=n\delta$ has probability $q(n)$ given by $q(n)n\delta = \sum_{i=1}^{h/\delta} q(n-i)\mu(i)i\delta$; in which $\mu(i) = \lambda T(i)$, with $T(i)$ being the total duration of the pulse with amplitude $h(t)$ satisfying $i\delta \le h(t) \le (i+1)\delta$. This formula could be used to compute $q(n)$ using the computed $q(n-i)$. We have to properly renormalize the probabilities afterwards.

If a pulse has both positive and negative values such as the monocycle pulses, we separate the pulses into two pulses, one having all positive values and the other having all negative values. Subsequently we can compute the amplitude distribution for say the positive pulses, as well as the amplitude distribution for the negative pulses. The overall amplitude distribution for the monocycle pulse is given by a convolution of the amplitude distributions.

We can now compute the probability of false detection of a pattern. A pattern is falsely detected if at the pattern sample time, all k taps each detects a pulse. The probability of false detection of a pattern is probability $[\Pr(z \ge h)]^k$. It has been shown in an earlier paper that the capacity of a single user is proportional to the logarithm of probability of false detection. Therefore as the number of tap k increases, the capacity increases linearly as k, if we keep λ constant. Thus if we increase the number of receivers, the capacity increases linearly.

4. Computing MAC Capacity

We model the channel as follows. Each user sends independently at the beginning of each unit time a pulse with probability p. With match filtering at the receiver, the output y is sampled. The MAI noise perturbs the value $x=0$ or 1 sent. Thus we can model the channel as a discrete time memoryless channel with known input probability $p(x)$ and channel transition probability density $p(y|x)$ as computed in the previous section. It should be noted that consecutive samples of the MAI noise are not independent and therefore the channel really is not memoryless. However, it is well known in Information Theory that channel memory can only increase channel capacity compared to the case with no memory, and that channel memory can be rendered memoryless by codeword interleaving.

We use the Shannon capacity formula to compute the capacity per user with one receive antenna. We then find the sum of capacity of N users, while keeping $Np=\lambda$ a constant. The MAC capacity various pulse shapes of figure 3 are shown respectively in figure 5-7.

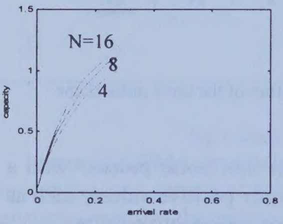


fig 5. Capacity -rectangular pulse

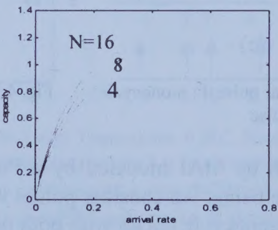


fig 6. Capacity-monocycle Gaussian pulse

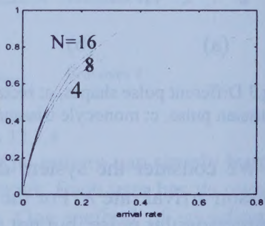


fig 7. Capacity -monocycle rect. Pulse

In conclusion, we have shown a large throughput for single antenna, and that capacity increases linearly as the number of receivers. The size of the receiver antenna array has to be of a dimension being multiples of the physical length of the pulse.

References:

1. J. Hui, "Pattern Code Modulation and Optical Decoding - A Novel Code-Division ultiplexing Technique for Multifiber Networks", IEEE JSAC,Nov. 1985.
2. M. Z. Win, R.A. Scholtz, "Impulse Radio: How it Works," *IEEE Comm. Letters,* Feb. 1998
3. F.Ramirez-Mireles, R.A. Scholtz, "Multiple Access Performance of Ultra-Wideband Spread Spectrum Impulse Radio Using Block Waveform Modulation," *IEEE ICC, June 1998.*
4. J. Hui, Spatial CDMA- A Distributed Multiple Antenna Access Method for Broadband Wireless Communication, *Proceedings of IEEE VTC.,* Boston, September 2000.

FACTORS INFLUENCING THE SHANNON CAPACITY OF SPACE TIME CODE

Joseph Y. Hui, Chunyu Bi, Hongxia Sun
Department of Electrical Engineering
Arizona State University, AZ 85287
(jhui@asu.edu)

Abstract

We study directly the geometric properties of electromagnetic waves to derive the channel capacity of multiple antenna systems. Spatial system transfer functions and spatial sampling theory are used, instead of the usual independent and complex Gaussian assumption.

1. Introduction

For space-time codes, it has been shown that communication capacity grows linearly as the number n of transmitters and receivers, under the assumption that each transmitter-receiver pair are identically and independently Raleigh faded [1-4]. In other words, h_{ij} are i.i.d. complex Gaussian random variables for all transmitter-receiver pair *i-j.* The validity of this independence assumption should be carefully examined, and we propose to approach the problem directly from EM wave equations.

To motivate the study, we consider the geometry shown in figure 1. We have the transmitting and receiving arrays being broadside and separated at a distance of D. The arrays each has n antennas with adjacent antennas separated by a distance of d. Figure 2 plots the MAE capacity formula if we only have direct line of sight. We observe that capacity reaches its first maximum when $nd^2 = D\lambda$, and thereafter oscillates with a constant mean plotted in figure 3 as a function of n. In fact for large enough d, capacity grows linearly as n, similar to that with the independence assumption. We also studied the multpath fading environment when we have two reflected rays, showing that the required d for a saturating capacity, though somewhat decreased, is still quite large. Hence to achieve a linear capacity increase while keeping d small, we must have dense multipaths.

While dense multipaths render h_{ij} truly independent, they also create delay spread and flat fading. Implicitly, we require a large symbol duration to avoid intersymbol interference. The linear increase in capacity in the spatial dimension is achieved with a reduction in capacity in the temporal dimension. There seems to be an intricate relationship between spatial frequencies and temporal frequency that is not well understood. If we examine the Helmholtz equation for wave propagation $\frac{\partial^2 E}{\partial x^2}+\frac{\partial^2 E}{\partial y^2}+\frac{\partial^2 E}{\partial z^2}-\frac{1}{c^2}\frac{\partial^2 E}{\partial t^2}=0$, a Fourier transform of this equation would yield the conservation relationship $f_x^{\,2}+f_y^{\,2}+f_z^{\,2}=\frac{1}{c^2}f_t^{\,2}$ within a volume. We shall show how this affects communication capacity.

2. System Function of Spatial Communications based on Helmholtz Equation.

The communication geometry of free space is shown in figure 4. The source *U,* system transfer function *H,* and sink *S* may be functions of spatial parameters *x, y, z,* f_x, f_y and temporal parameters *t,* f_t, which are selected in context. The source could be in discrete form if we transmit information via discrete antennas, hence $U(x,y)=\sum_{m,n} A_{mn}\delta(x-m\Delta_x, y-n\Delta_y)$. Reception is discrete, and we ignore the effects of antenna coupling. Typically, *U* and *S* have finite support Σ_0 and Σ_z in the respective planes *z*=0 and *z*, over which the Fourier transforms are obtained. Through techniques of Fourier optics [5], the linearity of the Helmholtz equation yields the following system transfer equation: $S(f_x,f_y,f_t)=U(f_x,f_y,f_t)H(f_x,f_y,f_t)$, where the system transfer function is

$$H(f_x,f_y,f_t)=\exp(j2\pi\frac{z}{c}\sqrt{f_t^2-(cf_x)^2-(cf_y)^2})$$

H suffers a rapid exponential decay if $f_t^2<(cf_x)^2+(cf_y)^2$. Thus free space limits the transfer of spatial frequencies to less than 1/λ. The transmitted signal *U* by itself should be band-limited because Δ_x and Δ_y should be larger than λ. The finite support Σ_0 somewhat broadens the spatial frequency. The finite support Σ_z distorts the reconstruction of the transmitted *U.*

A sampling theory can be derived from this formulation. The objective is to sample Σ_z such that we can recover the transmitted signals as much as possible. This objective is examined next.

3. Shannon Capacity in the Spatial Domain

A noise model is necessary for deriving the Shannon capacity given by the system transfer function. We assume $S(f_x,f_y,f_t)=U(f_x,f_y,f_t)H(f_x,f_y,f_t)+N(f_x,f_y,f_t)$,in which $N(f_x,f_y,f_t)$ is a Gaussian noise that is white in the spatial and temporal frequencies. We can inverse filter the received signal, subsequently obtaining the colored noise *N/H.* The Shannon capacity is therefore given by the water-filling formula [6] $C=\int\int\int\frac{1}{2}\log(1+\frac{(\nu-N/H(f_x,f_y,f_t))^+}{N/H(f_x,f_y,f_t)})df_x df_y df_t$ where the water level ν is chosen so that $\int\int\int(\nu-N/H(f_x,f_y,f_z))^+ df_x df_y df_z=P$, the transmitted power. We are now in a position to interpret how various factors may affect spatial capacity. It is noteworthy that the amplitude of $H(f_x,f_y,f_t)$ is 1 inside the region where the argument of the square root is real, i.e. $f_t^2>(cf_x)^2+(cf_y)^2$ for given f_t. The passband is shown in figure 5 (between shaded circles).

This passband allows the signal *U* through, so we should examine the distribution of *U.* If *U* is centered around a particular (f_x,f_y,f_t), more energy is passed if the center is close to the center of the disc shown, and if the energy around the center is well focused. These two conditions are obtained respectively by placing the transmitting antenna far apart, as well as having a Σ_0 that is large. This is congruent with our previous computation of channel capacity.

4. Conclusion

In a full paper, we shall fully explore how the transmission and reception should be better performed, and shed insight into the precise formulation of signal sampling in time and space. Also, we shall use this theory to examine explain how multipath could result in a higher channel capacity when the antennas are not far apart. Computation of channel capacities will be shown.

References:

1. E. Telatar, "Capacity of Multi-Antenna Gaussian Channels", *AT& T-Bell Labs Internal Tech. Memo.*, June 1995.
2. G. J. Foschini, Jr. and M. J. Gans, "On Limits of Wireless Communication in a Fading Environment When Using Multiple Antennas", *Wireless Personal Commun.*, Mar.98.
3. G. J. Foschini, Jr., "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multielement Antennas", *Bell System Technical Journal*, Autumn, 1996
4. V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-Time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction", *IEEE Transactions on Information Theory*, pp. 744-765, Max. 1998.
5. J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, 1968.
6. T. M. Cover, J. A. Thomas, *Elements of Information Theory,* Wiley Interscience, 1991.

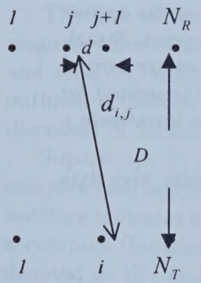


Fig 1. Communication geometry (D=1000λ, SNR=10dB)

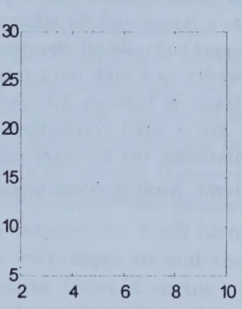


Fig 3. Average channel capacity versus *n*

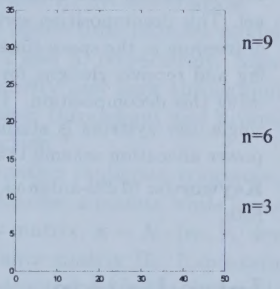


Fig 2. Channel capacity versus distance of antennas

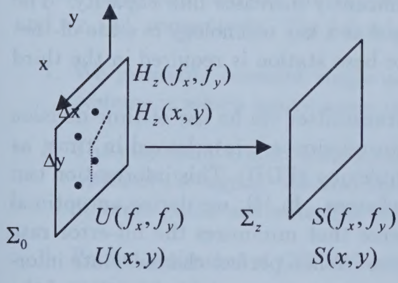


Fig 4. System transfer function for given frequency f_t

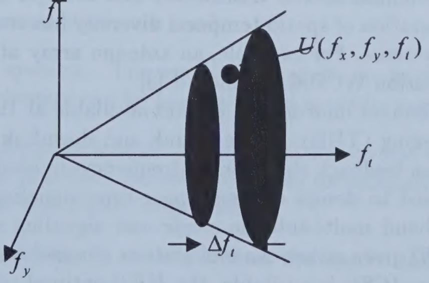


Fig 5. Passband of channel

Efficient Use of Wideband Spatio-temporal Wireless Channels Given Channel State Information*

Eko N. Onggosanusi, Barry D. Van Veen, and Akbar M. Sayeed

Department of Electrical and Computer Engineering
University of Wisconsin–Madison
Madison, WI 53706
`onggo@cae.wisc.edu, [vanveen,akbar]@engr.wisc.edu`

Abstract

A orthogonal decomposition of a general wideband space-time multipath channel is derived assuming antenna arrays at both the transmitter and receiver. Knowledge of channel state information at the transmitter is used to decompose the channel. This decomposition serves as a framework for efficiently managing the degrees of freedom in the space-time channel to trade-off diversity and throughput. Signaling and receiver designs for diversity and high data rate applications are derived using this decomposition. The trade-off between bit-error rate and throughput for single-user systems is studied. For a fixed throughput system, we investigate a power allocation scheme that minimizes the bit-error rate.

Keywords: Multi-antenna systems, wireless communications, diversity, high data rate.

1 Introduction

Use of antenna arrays at both the base station and mobile is envisioned in future wireless communication systems [1]. It is well-known (see e.g. [2]) that availability of multiple antennas at the transmitter and receiver significantly increases link capacity. The exploitation of spatio-temporal diversity has emerged as a key technology in state-of-the-art systems. For example, an antenna array at the base station is required in the third generation WCDMA standard [3].

Channel information is often available at the transmitter via he use of time-division duplexing (TDD), where uplink and downlink transmission are interleaved in time, as well as feedback channels for frequency-division duplexing (FDD). This information can be used to design efficient space-time signaling schemes. In [4], we derive an optimal wideband multi-antenna single-user signaling scheme that minimizes the bit-error rate (*BER*) given either channel *state* or channel *statistics*. When perfect channel state information (CSI) is available, the *BER*-optimal scheme transmits a weighted version of the same signature code in all antennas with the weights determined by the channel state. For a multipath channel, the optimal code is a discrete-time sinusoid with frequency equal to the space-time channel frequency having maximum gain. This optimal scheme can

---

*This research is supported in part by NSF Grant No. ECS-9979448.

be viewed as a generalization of selection diversity via adaptive frequency hopping with spatial beamforming. In this scheme, only one channel dimension is used in a particular symbol duration.

In this paper, the results in [4] are extended to high throughput applications. An orthogonal decomposition of the general L-path wideband space-time channel consisting of P-transmit and Q-receive antennas is derived. The channel modes are shown to be discrete-time sinusoids, (representing temporal dimensions) concatenated with spatial beamformers (representing spatial dimensions). This decomposition serves as a framework for efficiently managing the degrees of freedom in the space-time channel for any combination of diversity and high throughput applications. Maximum throughput is achieved by using all the channel dimensions as *parallel sub-channels* to transmit distinct data streams. Related ideas for high throughput transmission in narrow band systems are found in [5].

We assume that the same modulation scheme is used for all the sub-channels. The lowest BER is achieved by transmitting only via the sub-channel with the highest gain. Thus, this scheme has the lowest throughput. On the other hand, utilizing all sub-channels for maximizing the throughput results in the high BER in many sub-channels. We study the trade-off between BER and throughput by deriving a power allocation scheme that minimizes the BER for a fixed throughput and BPSK modulation.

The rest of the paper is organized as follows. The space-time channel model and singular value decomposition are described in Sections 2 and 3, respectively. Signaling and receiver designs for single-user with maximum diversity and high throughput are outlined in Section 4 and 5, respectively. Trade-off between throughput and diversity is discussed in Section 6, followed by the conclusions in Section 7.

Superscript T, H, and $*$ indicate matrix transpose, matrix conjugate transpose, and complex conjugation, respectively. Uppercase boldface denotes a matrix while lowercase boldface indicates a vector. $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. $\mathbf{x} \sim \mathcal{N}_C[\mathbf{m}, \mathbf{R}]$ denotes a complex Gaussian vector $\mathbf{x}$ with mean $\mathbf{m}$ and covariance matrix $\mathbf{R}$. Expectation is denoted as E[.] and the Euclidean norm of vector $\mathbf{x}$ is denoted as $\|\mathbf{x}\|$. The symbol $\otimes$ denotes Kronecker product and vec($\mathbf{A}$) is formed by stacking the columns of matrix $\mathbf{A}$ into a vector [6, 7]. $\mathbf{e}_n$ is the column vector $[\ \mathbf{0}^T\ 1\ \mathbf{0}^T]^T$ with 1 located at the n-th row. The diagonal matrix generated by the vector $\mathbf{v}$ is denoted by $\mathbf{diag}\{\mathbf{v}\}$. The i-th largest eigenvalue and the corresponding eigenvector of matrix $\mathbf{A}$ is denoted by $\lambda_i[\mathbf{A}]$ and $\mathbf{ev}_i[\mathbf{A}]$, respectively. The following assumptions are used:

1. We primarily consider single-user systems. The results also apply to multiuser systems in which multi-access interference can be approximated as white noise.

2. The maximum transit time across the array is small compared to the inverse bandwidth of the signal. We also assume that the channel coefficients corresponding to different paths and antennas are not completely correlated.

3. Perfect estimates of CSI and multipath delays are available at the transmitter and receiver.

2 General Single-User Multi-antenna Framework

Consider a single-user system with P transmit and Q receive antennas. We assume that each component of the transmitted signal $\mathbf{x}(t) = [\ x_1(t)\ ,\ x_2(t)\ , \cdots,\ x_P(t)\ ]^T$ has duration T and two-sided essential bandwidth B. The transmitted signal $\mathbf{x}(t)$ propagates over an L-path P-input, Q-output fading channel with delays $\{\tau_1, \tau_2, \cdots, \tau_L\}$. Without loss of generality, assume $\tau_1 < \tau_2 < \cdots < \tau_L$. Consider a typical wideband multipath channel with $\tau_L \ll T$. Hence, intersymbol interference is negligible and the baseband signal at the receiver $\mathbf{r}(t) \in \mathbb{C}^Q$ within 1 symbol interval $0 \le t < T$ can be written as

$$\begin{aligned} \mathbf{r}(t) &= \begin{bmatrix} r_1(t) \\ \vdots \\ r_Q(t) \end{bmatrix} = \sum_{l=1}^{L} \begin{bmatrix} \mathbf{h}_{l1}^T \\ \vdots \\ \mathbf{h}_{lQ}^T \end{bmatrix} \mathbf{x}(t-\tau_l) + \mathbf{n}(t) \, , \\ \mathbf{h}_{lq} &= \begin{bmatrix} h_{lq1} & \cdots & h_{lqP} \end{bmatrix}^T \, . \end{aligned} \tag{1}$$

The channel coefficient h_{lqp} corresponds to the l-th path, q-th receive antenna, and p-th transmit antenna. The additive noise vector $\mathbf{n}(t)$ is assumed to be zero mean circularly symmetric complex Gaussian with $E[\mathbf{n}(t)\mathbf{n}^H(t')] = \sigma^2\delta(t-t')\mathbf{I}_Q$. The signal $r_q(t)$ at the q-th receive antenna in (1) can also be written as

$$\begin{aligned} r_q(t) &= \sum_{l=1}^{L} \mathbf{h}_{lq}^T \mathbf{x}(t-\tau_l) + n_q(t) \\ &= \begin{bmatrix} \mathbf{x}^T(t-\tau_1) & \cdots & \mathbf{x}^T(t-\tau_L) \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{h}_{1q} \\ \vdots \\ \mathbf{h}_{Lq} \end{bmatrix}}_{\mathbf{h}_q} + n_q(t) \, . \end{aligned} \tag{2}$$

We assume that the p-th antenna waveform $x_p(t)$ has the following form:

$$x_p(t) = \sum_{i=0}^{N-1} s_p[i]\omega(t - i/B) \, , \quad 0 \le t < T \, , \tag{3}$$

where $\omega(t)$ is the (unit-energy) chip waveform of duration $1/B$ and N is defined such that $N = TB$. Since $x_p(t)$ is essentially bandlimited to B, assuming τ_l to be an integer multiple of $1/B$ suffices [8]. That is, $\tau_l = d_l/B$, $d_l \in \{0, 1, \cdots, N-1\}$. We sample $r_q(t)$ at the rate $\frac{1}{B}$ to enable discrete-time processing without loss of information. Let

$$\begin{aligned} \mathbf{r}_q &\overset{def}{=} [r_q(0), r_q(1/B), \cdots, r_q((N-1)/B)]^T \, , \\ \mathbf{s}_p &\overset{def}{=} [s_p[0], s_p[1], \cdots, s_p[N-1]]^T \, , \\ \mathbf{S} &\overset{def}{=} [\, \mathbf{s}_1 \cdots \mathbf{s}_P \,] \, . \end{aligned} \tag{4}$$

Define $\mathbf{\Delta}_{d_l} \in \mathbb{C}^{N \times N}$ as the time-shift matrix corresponding to the path delay d_l. We assume the delay is cyclic, so that $\mathbf{\Delta}_{d_l}$ is circulant. For example, when $d_l = 1$ and $N = 3$, $\mathbf{\Delta}_{d_l} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$. While the actual delay corresponds to a linear shift, negligible error is introduced for sufficiently large N provided that $d_L \ll N$. Furthermore, if a *chip-level* cyclic prefix is introduced, the cyclic shift is exact [8]. By defining $\mathbf{\Delta} \overset{def}{=} [\mathbf{\Delta}_{d_1} \cdots \mathbf{\Delta}_{d_L}] \in \mathbb{C}^{N \times NL}$, $\mathbf{s} = \mathrm{vec}(\mathbf{S}) \in \mathbb{C}^{NP}$, and $\mathbf{H}_q \overset{def}{=} [\mathbf{h}_{1q} \cdots \mathbf{h}_{Lq}] \in \mathbb{C}^{P \times L}$ such that $\mathbf{h}_q = \mathrm{vec}(\mathbf{H}_q)$, from (2) – (3) we have

$$\mathbf{r}_q = \mathbf{\Delta}\left(\mathbf{I}_L \otimes \mathbf{S}\right)\mathbf{h}_q + \mathbf{n}_q = \underbrace{\mathbf{\Delta}\left(\mathbf{H}_q^T \otimes \mathbf{I}_N\right)}_{\mathcal{H}_q}\mathbf{s} + \mathbf{n}_q \, , \tag{5}$$

where the second equality follows by applying the identity $\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^T \otimes \mathbf{A})\mathrm{vec}(\mathbf{X})$ [6] twice, and $\mathbf{n}_q \sim \mathcal{N}_C\left[\mathbf{0}, \sigma^2 \mathbf{I}_N\right]$.

3 Space-Time Channel Decomposition

The overall MIMO space-time channel is $\mathcal{H} \overset{def}{=} \left[\begin{array}{ccc} \mathcal{H}_1^T & \cdots & \mathcal{H}_Q^T \end{array} \right]^T \in \mathbb{C}^{NQ \times NP}$. Hence, the number of available space-time dimensions N_{dim} is precisely the rank of $\mathcal{H}$. Since we assume that the channel coefficients $\{h_{lqp}\}$ are not completely correlated (Assumption 2 in Section 1), $N_{dim} = N \times \min(P, Q)$ w.p.1. Our goal is to design transceivers that access all N_{dim} degrees of freedom. Obviously, this goal can be accomplished when an SVD of $\mathcal{H}$ is available. This SVD has to be computed numerically since for $Q > 1$, a closed-form SVD for $\mathcal{H}$ can not be obtained. Numerical computation is prohibitive in practice since N is large for wideband applications (≥ 32).

In this paper, instead of using an SVD for $\mathcal{H}$, we derive a closed-form SVD for $\mathcal{H}_q$ and show that via appropriate signaling and receiver designs, all N_{dim} degrees of freedom can be accessed. The circulant structure of $\mathbf{\Delta}_{d_l}$ is exploited to obtain a singular value decomposition (SVD) of the q-th receive antenna space-time channel matrix $\mathcal{H}_q$ in (5).

Theorem 1 *$\mathcal{H}_q \in \mathbb{C}^{N \times NP}$ admits the following* SVD:

$$\begin{aligned}
\mathcal{H}_q &= \mathbf{F}\mathbf{\Sigma}_q \mathbf{\Upsilon}_q^H , &(6)\\
\mathbf{F} &= \left[\begin{array}{cccc} \mathbf{c}_0 & \mathbf{c}_1 & \cdots & \mathbf{c}_{N-1} \end{array} \right] , &(7)\\
\mathbf{\Sigma}_q &= \mathbf{diag}\left\{ \|\mathbf{g}_{0,q}\| , \|\mathbf{g}_{1,q}\| , \cdots \|\mathbf{g}_{N-1,q}\| \right\} , &(8)\\
\mathbf{\Upsilon}_q &= \left[\begin{array}{cccc} \frac{\mathbf{g}_{0,q}}{\|\mathbf{g}_{0,q}\|} \otimes \mathbf{c}_0 & \frac{\mathbf{g}_{1,q}}{\|\mathbf{g}_{1,q}\|} \otimes \mathbf{c}_1 & \cdots & \frac{\mathbf{g}_{N-1,q}}{\|\mathbf{g}_{N-1,q}\|} \otimes \mathbf{c}_{N-1} \end{array} \right] , &(9)\\
\mathbf{c}_n &\overset{def}{=} \frac{1}{\sqrt{N}} \left[1, e^{-j2\pi n/N}, \cdots, e^{-j2\pi(N-1)n/N} \right]^T , &(10)\\
\mathbf{g}_{n,q} &\overset{def}{=} \frac{1}{\sqrt{N}} \left[\begin{array}{c} \sum_{l=1}^L h_{lq1}^* e^{j2\pi n d_l/N} \\ \vdots \\ \sum_{l=1}^L h_{lqP}^* e^{j2\pi n d_l/N} \end{array} \right] . &(11)
\end{aligned}$$

Proof: Appendix.

Notice that the left singular unitary matrix $\mathbf{F}$ is the N-IDFT matrix, and is independent of q since it is associated with temporal channel characteristics. Also, notice that the p-th row of $\mathbf{g}_{n,q}$ is the complex conjugate of the frequency response of the channel between the q-th receive and p-th transmit antenna at frequency $\frac{2\pi}{N}$.

Consider front-end matched filtering with *only* the channel coefficients and coherent combining of the received signal from all received antennas as denoted by $\tilde{\mathbf{r}} \overset{def}{=} \sum_{q=1}^{Q} \mathcal{H}_q^H \mathbf{r}_q$. Then, by Theorem 1 and the identity $(\mathbf{X}_1 \otimes \mathbf{X}_2)(\mathbf{Y}_1 \otimes \mathbf{Y}_2) = \mathbf{X}_1\mathbf{Y}_1 \otimes \mathbf{X}_2\mathbf{Y}_2$, we have

$$\begin{aligned}
\tilde{\mathbf{r}} &= \left(\sum_{q=1}^{Q} \mathcal{H}_q^H \mathcal{H}_q \right) \mathbf{s} + \sum_{q=1}^{Q} \mathcal{H}_q^H \mathbf{n}_q &(12)\\
\sum_{q=1}^{Q} \mathcal{H}_q^H \mathcal{H}_q &= \sum_{n=0}^{N-1} \mathbf{\Gamma}_n \otimes \mathbf{c}_n \mathbf{c}_n^H \in \mathbb{C}^{PN \times PN} &(13)\\
\mathbf{\Gamma}_n &\overset{def}{=} \sum_{q=1}^{Q} \mathbf{g}_{n,q} \mathbf{g}_{n,q}^H \in \mathbb{C}^{P \times P} . &(14)
\end{aligned}$$

Since the rank of $\mathbf{\Gamma}_n$ in (14) is $\min(P,Q)$, it is clear from (13) that $\sum_{q=1}^{Q} \mathcal{H}_q^H \mathcal{H}_q$ is of rank $N_{dim} = N \times \min(P,Q)$ w.p.1. It is easy to verify that the eigenvectors of $\sum_{q=1}^{Q} \mathcal{H}_q^H \mathcal{H}_q$ are

$$\begin{aligned} &\mathbf{w}^{((i-1)N+n)} \otimes \mathbf{c}_n \ , \quad n \in \{0,1,\cdots,N-1\} \ , \quad i \in \{1,2,\cdots,\min(P,Q)\} \\ &\mathbf{w}^{((i-1)N+n)} = \mathbf{ev}_i\left[\mathbf{\Gamma}_n\right] \ . \end{aligned} \tag{15}$$

with the corresponding eigenvalues $\gamma_{(i-1)N+n+1} = \lambda_i\left[\mathbf{\Gamma}_n\right]$. These N_{dim} eigenmodes represent all the available *orthogonal space-time sub-channels.*

Notice that to compute all the eigenvectors in (15), the most costly operation is finding the eigenvectors of N matrices, each with the size of $P \times P$. This is much less complex compared to computing the singular modes of $\mathcal{H}$ which is $NQ \times NP$ since N is typically large (≥ 32) and P is typically small (currently 2) in practice.

4 Single-User with Maximum Diversity

Maximum (selection) diversity gain is obtained by transmitting only via the most dominant sub-channel [4]. In this case, we choose

$$\mathbf{s} = \sqrt{\rho}\, b \left(\mathbf{w} \otimes \mathbf{c}_{\overline{n}}\right) \tag{16}$$

where b and ρ are the data symbol and power, respectively, and

$$\mathbf{w} = \mathbf{ev}_1\left[\mathbf{\Gamma}_{\overline{n}}\right] \ , \quad \overline{n} = \arg \max_{n=0,\cdots,N-1} \lambda_1\left[\mathbf{\Gamma}_n\right] \ . \tag{17}$$

That is, $\overline{n}$ corresponds to the $\mathbf{\Gamma}_n$ with the largest dominant eigenvalue and $\mathbf{w}$ is the corresponding the dominant eigenvector of $\mathbf{\Gamma}_n$. Notice that only *one* dimension is used in one symbol duration to achieve maximum diversity gain.

For simplicity, we assume BPSK modulation ($b \in \{\pm 1\}$) for receiver design although extension to other modulation schemes is straightforward. At the receiver, $\tilde{\mathbf{r}}$ in (12) is computed and then correlated with $\mathbf{w} \otimes \mathbf{c}_{\overline{n}}$ to obtain test statistic for b. To further reduce the complexity, we exploit Theorem 1 as follows. It is easy to show using the identity $(\mathbf{X}_1 \otimes \mathbf{X}_2)(\mathbf{Y}_1 \otimes \mathbf{Y}_2) = \mathbf{X}_1\mathbf{Y}_1 \otimes \mathbf{X}_2\mathbf{Y}_2$ and the orthogonality of $\{\mathbf{c}_n\}_{n=0}^{N-1}$ that $\mathbf{\Sigma}_q \mathbf{\Upsilon}_q^H \mathbf{s} = \left(\mathbf{g}_{\overline{n},q}^H \mathbf{w}\right) \mathbf{e}_{\overline{n}}$. Hence

$$\tilde{\mathbf{r}}_q = \sqrt{\rho}\, b \left(\mathbf{g}_{\overline{n},q}^H \mathbf{w}\right) \mathbf{c}_{\overline{n}} + \mathbf{n}_q \ .$$

Hence, computing $\tilde{\mathbf{r}}$ and correlating with $\mathbf{w} \otimes \mathbf{c}_{\overline{n}}$ can be done as shown in in Figure 1 with $n = \overline{n}$. Note that $\{\mathbf{g}_{n,q}\}$, which represents the CSI, and the transmit beamformer $\mathbf{w}$ need to be known at the receiver.

5 High Throughput Single-User Systems

Since we are assuming a fixed modulation scheme, the throughput of the system is determined by the number of streams transmitted simultaneously. To transmit M data streams via the channel, we choose

$$\mathbf{s} = \sum_{m=1}^{M} \sqrt{\rho_m}\, b_m\, \mathbf{s}^{(m)} \ , \tag{18}$$

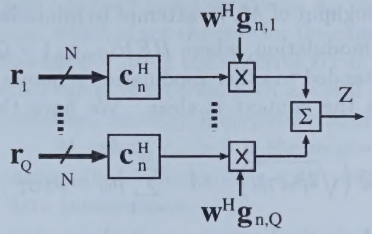


Figure 1: Receiver structure for single-user system with maximum diversity.

where the signature codes $\{s^{(m)}\}$ are chosen to be a subset of the eigenvectors in (15). Define the *relative throughput* of a system as the total system throughput relative to the throughput of a one-stream system with the same modulation scheme. The relative throughput M is bounded by N_{dim}, the maximum number of parallel sub-channels.

At the receiver, different streams can be separated due to the orthogonality of $\{s^{(m)}\}$. The receiver for this maximum throughput scheme may be implemented using the front-end structure in Figure 1 for each data stream with n and $\mathbf{w}$ chosen accordingly.

6 Trade-off between Diversity and Throughput

For a given modulation scheme, throughput is maximized by using all the N_{dim} spatio-temporal dimensions for data transmission as discussed above. However, for a fixed total transmitted power, this comes at the expense of BER since the power has to be distributed between N_{dim} streams. Hence, there is a trade-off between throughput and BER. To attain lower BER for high throughput systems, channel coding is usually employed at the cost of reduced throughput.

Instead of maximizing system throughput, one may trade throughput for lower BER by choosing a scheme with $M < N_{dim}$. This trade-off between throughput and BER may be studied using the framework of Theorem 1. Since perfect CSI is available at the transmitter, the sub-channel SNR values (eigenvalues) $\{\gamma_m\}_{m=1}^{N_{dim}}$ can be computed. Without loss of generality, assume that

$$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_M \geq \gamma_{M+1} \geq \cdots \geq \gamma_{N_{dim}} . \tag{19}$$

Clearly, the most power-efficient way to achieve a relative throughput of M is to use the M sub-channels with the highest SNR values. We define the *effective* BER of the M-stream system as follows:

$$BER_{eff}^{(M)} \stackrel{def}{=} \frac{1}{M} \sum_{m=1}^{M} BER(\rho_m \gamma_m) . \tag{20}$$

where $BER(\rho_m \gamma_m)$ is the BER corresponding to the m-th stream, which is a strictly decreasing function of the received SNR $\rho_m \gamma_m$ and depends on the chosen modulation scheme. The transmit power allocated for all streams $\{\rho_m\}_{m=1}^{M}$ must satisfy the constraint $\sum_{m=1}^{M} \rho_m = \rho_{TOT}$. In general, given a fixed total power ρ_{TOT}, $BER_{eff}^{(M)}$ increases as M is increased. This concept for trading BER and throughput can be applied in practical wireless systems to dynamically satisfy changing throughput and BER requirements. We now consider a strategy for optimizing this trade-off.

For a given relative throughput of M, we attempt to minimize $BER_{eff}^{(M)}$ with respect to $\{\rho_m\}_{m=1}^{M}$ assuming BPSK modulation, where $BER(\rho_m\gamma_m) = \mathcal{Q}\left(\sqrt{2\rho_m\gamma_m}\right)$. The results presented below can be extended to other modulation schemes. For clarity, we suppress the superscript (M) when the context is clear. We have the following optimization problem:

$$\min_{\rho_1,\cdots,\rho_M} \sum_{m=1}^{M} \mathcal{Q}\left(\sqrt{2\rho_m\gamma_m}\right) \quad \text{s.t.} \quad \sum_{m=1}^{M} \rho_m = \rho_{TOT}\ , \quad \rho_m \geq 0\ . \tag{21}$$

This problem can be solved via the Lagrange multiplier technique. It can be shown that the optimality conditions can be written as

$$\sqrt{\frac{\gamma_m}{\rho_m}} \times \exp\left(-\rho_m\gamma_m\right) = \mu\ , \quad m = 1,\cdots,M \tag{22}$$

$$\sum_{m=1}^{M} \rho_m = \rho_{TOT}\ . \tag{23}$$

where μ is chosen to satisfy the power constraint (23). The solution of (22) can be obtained numerically.

An approximate solution of (21) can be obtained by substituting the exact BER of each sub-channel in (21) with its Chernoff bound. Hence, we can also minimize the following upper bound

$$BER_{eff}^{(M)} \leq \frac{1}{2M} \sum_{m=1}^{M} \exp\left(-\rho_m\gamma_m\right)\ .$$

This approximation is accurate for sufficiently large ρ_{TOT}. Using Kuhn-Tucker condition, we obtain the following closed-form solution for ρ_m, $m = 1,\cdots,M$:

$$\rho_m = \max\left(0, \frac{\log\gamma_m - \mu}{\gamma_m}\right) \tag{24}$$

where μ is chosen to satisfy the power constraint in (23). We term this solution the *Chernoff-based* power allocation.

To get some insight on the minimum effective BER power allocation (based on the exact and Chernoff-based effective BER) we compare them to the uniform power allocation ($\rho_m = \rho_{TOT}/M$) and waterfilling ($\rho_m = \max(0, \mu - \gamma_m^{-1})$). We consider one channel realization for a $P = Q = 2$, $L = 3$ system with $d_l \in \{0,1,2\}$, $N = 16$, and $M = 32$. We assume that the channel coefficients $\{h_{lqp}\}$ are IID and $\mathcal{N}_C\,[0, 1/QL]$. The resulting sorted sub-channel SNR values are depicted in Figure 2. The power allocation and BER across sub-channels for $\rho_{TOT} = 20$ and 35 dB are depicted in Figure 3 (a)–(d). From this example, we make three observations:

1. The Chernoff-bound-based solution deviates from the exact solution for small ρ_{TOT} and approaches the exact solution as ρ_{TOT} increases.

2. For low ρ_{TOT}, the exact solution allocates more power to sub-channels with high SNR values. This is similar in principle to a waterfilling solution. However, the opposite holds for high ρ_{TOT}. For moderate ρ_{TOT}, more power tends to be allocated to sub-channels with moderate SNR values.

3. For the waterfilling and Chernoff-based solutions, ρ_m can be zero when γ_m is very small compared to ρ_{TOT}, which is not the case for the exact minimum effective BER solution. This is a drawback of using the effective BER criterion in (21) since sub-channels with very small values of received SNR are still used for data transmission. To overcome this limitation, we may incorporate additional constraints such as $\rho_m \gamma_m \geq c_m$, $m = 1, \cdots, M$, where $c_m > 0$ is the received SNR floor for the m-th sub-channel. This ensures that sub-channels with very low received SNR values will not be used for data transmission.

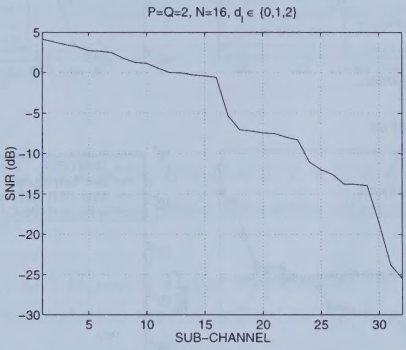


Figure 2: Sub-channel SNR values obtained from a realization of $N = 16$, $P = Q = 2$, $L = 3$ system with $d_l \in \{0, 1, 2\}$ and $M = 32$.

To compare the average performance of different power allocation schemes for various throughputs, we simulate the *average* BER_{eff} for $P = Q = 2$, $N = 16$ systems with $L = 3$ ($d_l \in \{0, 1, 2\}$) with the same channel statistics as the previous example. The simulation results are depicted in Figure 4 (a)–(b). Observe that the performance of Chernoff-based solution is very close to that of the exact minimum $BER_{eff}^{(M)}$ solution, especially for large ρ_{TOT}.

7 Conclusions

An orthogonal decomposition of a general space-time multi-antenna multipath channel is derived and utilized to design efficient signaling strategies and the corresponding receiver structures. The decomposition identifies $N \times \min(P, Q)$ available spatio-temporal dimensions in the channel. The factor N is the length of each signature code, which represents the number of available temporal degrees of freedom. The number of available spatial degrees of freedom is $\min(P, Q)$, where P and Q are the number of transmit and receive antennas, respectively. This decomposition serves as a framework to jointly address system design for maximum diversity and maximum throughput as well as the combination of the two, when channel state information (CSI) is available at the transmitter. We study the trade-off between BER, which is associated with diversity, and throughput for single user systems by separately investigating fixed throughput and fixed BER criteria. For a fixed throughput system, a power allocation scheme that minimizes the instantaneous BER for multistream system is derived.

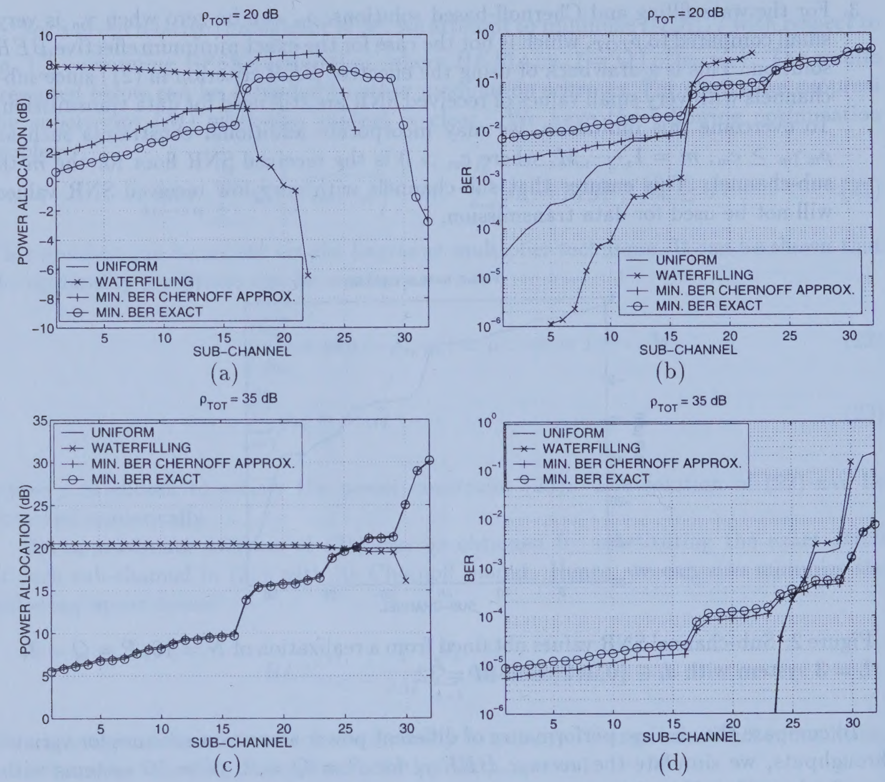


Figure 3: Different schemes of power allocation with the resulting sub-channel BER for $\rho_{TOT} = 20$ dB ((a)–(b)) and $\rho_{TOT} = 35$ dB ((c)–(d)).

Appendix: Proof of Theorem 1

Since $\mathbf{\Delta}_{d_l}$ is circulant, $\mathbf{\Delta}_{d_l} = \mathbf{F}\mathbf{\Lambda}_{d_l}\mathbf{F}^H$, where $\mathbf{F} \in \mathbb{C}^{N\times N}$ is the N-DFT matrix and [7]

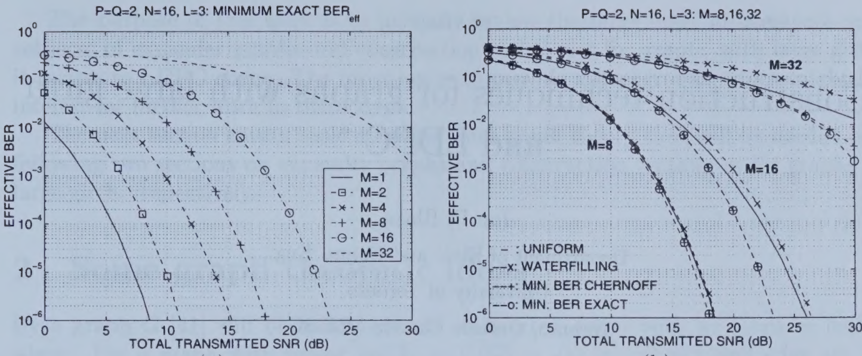
$$\mathbf{\Lambda}_{d_l} = \mathbf{diag}\left\{\mathbf{F}\mathbf{e}_{N+1-d_l}\right\} = \frac{1}{\sqrt{N}}\,\mathbf{diag}\left\{1, e^{j2\pi d_l/N}, \ldots, e^{j2\pi(N-1)d_l/N}\right\}.$$

Let $\mathbf{D}_{qp} \overset{def}{=} \sum_{l=1}^{L} h_{lqp}\mathbf{\Lambda}_{d_l}$. Then,

$$\begin{aligned}\mathcal{H}_q &= \mathbf{\Delta}\left(\mathbf{H}_q^T \otimes \mathbf{I}_N\right) = \left[\ \sum_{l=1}^{L} h_{lq1}\mathbf{\Delta}_{d_l} \ \cdots \ \sum_{l=1}^{L} h_{lqP}\mathbf{\Delta}_{d_l}\ \right] \\ &= \mathbf{F}\left[\,\mathbf{D}_{q1} \cdots \mathbf{D}_{qP}\,\right]\left(\mathbf{I}_P \otimes \mathbf{F}\right)^{H}. \end{aligned} \tag{25}$$

Using the same argument as in [4], it can be shown that

$$\left[\,\mathbf{D}_{q1} \cdots \mathbf{D}_{qP}\,\right] = \begin{bmatrix} \|\mathbf{g}_{1,q}\| & & \\ & \ddots & \\ & & \|\mathbf{g}_{N,q}\| \end{bmatrix} \begin{bmatrix} \mathbf{g}_{1,q}^H/\|\mathbf{g}_{1,q}\| & & \\ & \ddots & \\ & & \mathbf{g}_{N,q}^H/\|\mathbf{g}_{N,q}\| \end{bmatrix} \mathbf{\Pi}_{(P,N)}^H, \tag{26}$$



(a) (b)

Figure 4: Averaged $BER_{eff}^{(M)}$ for $N = 16$, $P = Q = 2$, $L = 3$ ($d_l \in \{0, 1, 2\}$) with different M. (a) Minimum $BER_{eff}^{(M)}$ (b) different power allocation schemes

where $\mathbf{g}_{N,q}$ is given in (11) and $\mathbf{\Pi}_{(P,N)} = \mathbf{\Pi}_{(P,N)} = \left[\mathbf{I}_P \otimes \mathbf{e}_1 \quad \mathbf{I}_P \otimes \mathbf{e}_2 \quad \cdots \quad \mathbf{I}_P \otimes \mathbf{e}_N \right] \in \mathbb{C}^P$ is the (P, N) unitary permutation matrix defined in [6]. Hence,

$$(\mathbf{I}_P \otimes \mathbf{F})\, \mathbf{\Pi}_{(P,N)} = \left[\mathbf{I}_P \otimes \mathbf{F}\mathbf{e}_1 \quad \cdots \quad \mathbf{I}_P \otimes \mathbf{F}\mathbf{e}_N \right] = \left[\mathbf{I}_P \otimes \mathbf{c}_0 \quad \cdots \quad \mathbf{I}_P \otimes \mathbf{c}_{N-1} \right] , \quad (27)$$

where $\mathbf{c}_n$ is defined in (10) and the first equality follows from the identity

$$(\mathbf{X}_1 \otimes \mathbf{X}_2)(\mathbf{Y}_1 \otimes \mathbf{Y}_2) = \mathbf{X}_1\mathbf{Y}_1 \otimes \mathbf{X}_2\mathbf{Y}_2 \,, \quad (28)$$

with appropriate dimensions. The proof is completed by combining (25), (26), and (27).

References

[1] J. H. Winters, "Smart antennas for wireless systems," *IEEE Personal Communications*, pp. 23–27, February 1998.

[2] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *AT & T Technical Memorandum*, 1995.

[3] *http://www.3gpp.org.*

[4] E. N. Onggosanusi, B. D. V. Veen, and A. M. Sayeed, "Optimal antenna diversity signaling for wideband systems utilizing channel side information," *submitted to IEEE Trans. Commun.*, 2000.

[5] G. G. Raleigh and J. M. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, pp. 357–366, March 1998.

[6] J. Brewer, "Kronecker products and matrix calculus in system theory," *IEEE Trans. Circ. System*, vol. CAS-25, pp. 772–781, Sept. 1978.

[7] G. H. Golub and C. F. V. Loan, *Matrix Computations.* John Hopkins, 1996.

[8] J. G. Proakis, *Digital Communications.* New York: McGraw Hill, 3rd ed., 1995.

Construction techniques for graphs with large girth and LDPC *

Ian F. Blake
Department of Elec. and Comp. Eng.,
University of Toronto,
Toronto, Ontario Canada M5S 3G4
ifblake@comm.utoronto.ca

Abstract

The study of bipartite graphs in the representation and analysis of low density parity check codes, as initiated by the work of Tanner [37], has received considerable attention in recent years. The use of iterative decoding algorithms on such graphs has been shown to be capable of achieving near-Shannon limit performance. This work partially reviews the recent graph-theoretic literature in the areas of expander graphs and construction techniques for graphs with large girth in the hope it might find application to the design of good LDPC codes and their encoding/decoding algorithms and analysis.

1 Introduction

Interest in iterative decoding algorithms, originally introduced by Elias [10] for product codes and Gallager [13] for low density parity check codes, has been rekindled by the remarkable performance achievable by turbo codes. The paper of Tanner [36] introduced a graphical approach for representing the low-density parity check matrix and interpreting the interactive decoding algorithms on them and this approach has proved influential in much of the current literature.

Recent analytical and experimental investigations on these codes show that near-Shannon limit error performance can be achieved with such iterative decoding algorithms by the appropriate design of the constraint or Tanner graph. Such studies indicate that it is not so much the minimum distance of the code that is the major factor of code performance as it is the match of the graph properties with the needs of the iterative decoder. Thus low vertex degrees and few short cycles, as well as a reasonable code minimum distance for the given code rate and length, seem important parameters in determining code performance. The recent work of Richardson et al [33], indicate that excellent error performance can be achieved with irregular bipartite graphs by choosing the degree sequences carefully. In a related direction, the work of Sipser and Spielman [35] shows that excellent performance can be achieved by using expander graphs. There remain many unanswered questions and these are clearly important directions to explore further.

---

*This research supported by NSERC grant number 7382

The purpose of this work is to partially review the large body of literature on the subjects of expander graphs and construction techniques for graphs with large girth as it relates to LDPC codes. The intention is to provide a background which might be of interest for further work in these areas.

The next section covers basic concepts from graph theory that will be needed in the following two sections on expander graphs and construction techniques for graphs with large girth, respectively.

2 Some graph theoretic ideas

By a graph G [41] will be meant a finite undirected graph with no loops or multiple edges. For a graph with vertex set $V = V(G) = \{v_1, v_2, \cdots, v_n\}$ and edge set $E = E(G) = \{e_1, e_2, \cdots, e_m\}$, we say G has order $n = |V|$ and size $m = |E|$ and vertices $u,\ v \in V$ are adjacent in G or neighbors if $(u, v) \in E$. The graph G is bipartite if its vertex set is the union of two disjoint sets $V = V_1 \cup V_2$ such that each edge joins a vertex from each set. A path in G is an ordered list of disjoint vertices $v_1, \cdots, v_k$ such that $(v_{i-1}, v_i) \in E$ for $i = 2, \cdots k$ and a cycle is a path for which $(v_k, v_1) \in E$. The graph is connected if there is a path between any two vertices. The girth of a graph is the length of a shortest cycle and the diameter is the maximum of distances between any two vertices.

The adjacency matrix of G of order n is $A = (a_{ij}),\ a_{ij} = 1$ if (v_i, v_j) in E, a real symmetric $n \times n$ matrix with zeros on the main diagonal and with real eigenvalues. The degree of a vertex is the number of edges incident with it and if all vertices have the same degree k the graph is referred to as k-regular. More generally, let $\delta(G)$ be the minimum valency of the edges of a graph and $\Delta(G)$ the maximum valency and G is k-regular if $\delta(G) = \Delta(G) = k$.

The number of vertices in a regular graph with valency k and odd girth $g = 2d+1$ is at least [5]

$$n_0(k, g) = 1 + k + k(k-1) + \cdots + k(k-1)^{(g-3)/2} = \frac{k(k-1)^d - 2}{k-2}$$

and with even girth $g = 2d$ is at least

$$n_o(k, g) = 1 + k + k(k-1) + \cdots + k(k-1)^{(g-4)/2} + (k-1)^{(g-2)/2} = \frac{2(k-1)^d - 2}{k-2}.$$

A (k, g)-graph is any k-valent graph with girth g and one with order achieving this lower bound, $n_0(k, g)$, will be referred to as a (k, g)-cage [42, 30, 23, 34]. Graphs which achieve these lower bounds tend to have remarkable properties [23]. Some properties of cubic graphs have been noted on the web site [34].

The problem of upper bounds on (k, g)-graphs tends to be a little harder. The strongest results currently known are due to Lazebnik et al [23], namely:

Theorem:[23] Let $k \geq 2$ and $g \geq 5$ be positive integers and let q denote the smallest odd prime power for which $k \leq q$. Then

$$v(k, g) \leq 2kq^{\frac{3}{4}g-a}$$

where $a = 4,\ 11/4,\ 7/2,\ 13/4$ for $g \equiv 0,\ 1,\ 2,\ 3$ mod 4 respectively.

Constructions which approach these bounds are noted in [23].

For a fixed integer $k \geq 3$ let $\{G_i\}_{i \geq 1}$ be a family of k-regular graphs of increasing order n_i with girth g_i. The lower bounds on $n(k, g)$ can be expressed [5] as

$$n_i(g) > A(k-1)^{g_i/2} + B$$

where A and B are constants independent of g. Consequently we say that a *family of graphs has large girth* if

$$g_i \geq \gamma \log_{k-1}(n_i)$$

for some constant γ where the above argument shows $\gamma \leq 2$. The largest constant known for families of graphs is $4/3$ and one such class will be discussed below.

The eigenvalues of the adjacency matrix of a graph are of interest. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues, including multiplicities, of the adjacency matrix. For a k-regular graph one has $\lambda_1 = k$ and $|\lambda_i| \leq k$ for all i [5]. Indeed, the smallest eigenvalue of a k-regular graph is $\lambda_n \geq k - n$ and in the case of bipartite regular, it is $-k$. The largest eigenvalue, particularly for regular graphs, is not of interest. It turns out that the second largest eigenvalue in absolute value, is of great interest with a large body of literature concerned with it. The problem will be pursued in the next section. We mention the influential result of Tanner [36] that seems to have inspired this approach. For a bipartite graph G with $|V_1| = n$ and $|V_2| = m$ and degrees $r+1$ and $s+1$ respectively, let M be its $n \times m$ incidence matrix and MM^T a matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Then a *bounded strong concentrator* (bsc) is a graph such that if $X \subset V_1$ such that $|X| \leq \alpha n$ then $|N(X)| \geq c|X|$. Tanner showed that if $\lambda_1 > \lambda_2$ then G is a bsc with

$$c \geq \frac{(s+1)^2}{[\alpha((s+1)(r+1) - \lambda_2) + \lambda_2]}.$$

It follows that the lower the value of λ_2 the greater the magnification. This line of investigation has been very significant to the area and pursued over the past two decades.

Cayley graphs account for many but not all of the vertex transitive graphs. They are constructed from an abstract group, G with identity, 1. Let Ω be a set of generators for G such that

$$x \in \Omega \Rightarrow x^{-1} \in \Omega, \quad 1 \notin \Omega.$$

The the Cayley graph has vertex set $V(G) = G$ and edge set

$$E(G) = \{(g,h) | g^{-1}h \in \Omega\}.$$

In the context of coding, a bipartite graph G with vertex partition $V = V_1 \cup V_2$ with the vertices of V_1 referred to as the bit vertices [38] and those of V_2 the parity vertices, is called the Tanner graph of a code if the $|V_1| \times |V_2|$ matrix formed by the incidence of edges between the two sets is the parity check matrix of the code. A slightly more general setting was used in [37]. Here the parity vertices were m-valent and there is assumed a code C of length m (referred to as a subcode) with minimum distance d. Then a binary word of length $n = |V_1|$ is a codeword if the binary m-tuple corresponding to bit vertices incident with a given parity vertex, is a codeword in C. The minimum distance, D, of the overall code is then bounded by [37]:

$$D \geq \begin{cases} d\frac{[(d-1)(m-1)]^{(g-2)/4}-1}{(d-1)(m-1)-1} + \frac{d}{m}[(d-1)(m-1)]^{(g-2)/4}, & \text{for } g/2 \text{ odd} \\ d\frac{[(d-1)(m-1)]^{g/4}-1}{(d-1)(m-1)-1} & \text{for } g/2 \text{ odd} \end{cases}$$

While the bound is exponential in the girth it does not seem tight in most cases considered.

The more common usage of the graph is to use as the code C the simple parity code of length m - thus the bit vertices adjacent to a given parity vertex has digits that have even parity. Tanner [38] has also recently considered minimum distance bounds for these codes. If H is the parity check matrix obtained from the graph and $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_s$ are the distinct eigenvalues of $H^T H$ (which are real) then if the bit vertices are m-regular and the parity vertices are j-regular and $|V_1| = n,\ |V_2| = r$ then the minimum distance of the code satisfies the bounds:

$$d \geq \frac{n(2m - \mu_2)}{(mj - \mu_2)}$$

and

$$d \geq \frac{2n(2m + j - 2 - \mu_2)}{j(mj - \mu_2)}.$$

For coding, it will be important to construct families of irregular graphs [33] and deterministic results on this problem seem limited. A graph for which the degrees of the vertices of V_1 are r and those of V_2 are s and the graph has girth $2t$ is referred to as an (r, s, t) graph [12] and it is shown there that such graphs exist for all $r,\ s,\ t \geq 2$. Certain types of constructions are also reported there.

3 Expanders and the second eigenvalue problem

Brief comments on the use of expander graphs in coding theory and the second eigenvalue problem are considered in this section. We assume here that $G = (V, E),\ |V| = n, |E| = e$ is a finite, simple, connected and k-regular graph, not necessarily bipartite although that will be the case of most interest. For a subset $X \subseteq V$, define

$$N(X) = \{v \in V,\ v \notin X | v \text{ adjacent to some } u \in X\}.$$

Informally, an expander is a graph in which certain sets of vertices have a "large" set of neighbors. The formal definition varies slightly in the literature and we use those of [4, 41].

Definition: G is an (n, k, c) *magnifier* if for every subset $X \subseteq V, |N(X)| \geq c|X|$. The largest such c is called the magnification of G.

It can be easily shown that $0 < c < 1$ or $(n+1)/(n-1)$ depending on whether n is even or odd.

In the case that G is bipartite the concept of expanding graphs is important:

Definition: G is an (n, k, c) *expander* if it a balanced bipartite $(V_1 = V_2 = n)$ if for every subset $X \subset V_1,\ |X| \leq n/2$,

$$N(X) \geq |X|(1 + c(1 - \frac{|X|}{n})).$$

The largest such c will be called the expansion of G. Such expander graphs have a variety of applications in switching networks, sorting networks and in establishing space-time trade-offs for certain types of computational problems.

If the previous definition holds for all $X \subset V_1$ it will be called a *strong* expander.

Since the present paper is concerned with graph theory rather than coding theory directly, only a brief look at the relationship between the two is noted here, following the work of Sipser and Spielman [35], who actually used unbalanced magnifier graphs, by our definition. We note only the theorem 7 of [35]:

Theorem: [35] Let B be a (c,d)-biregular graph with magnification $(c/d\epsilon)$ on sets of size at most αn, and C a subcode as noted in the Tanner construction of the previous section, with block length d, rate $r > (c-1)/c$ and minimum distance ϵ. Then the overall code has rate at least $cr - (c-1)$ and minimum distance αn.

Sipser and Spielman [35] constructed unbalanced bipartite graphs $B(n,d)$ by the following construction. Let $G(n,d)$ be a d-regular expander on n vertices and notice it has $nd/2$ edges. The graph $B(n,d)$ is then obtained by identifying one vertex set V_1 with the edges of $G(n,d)$ and the other, V_2, with the vertices of $G(n,d)$. Two vertices between these two sets are joined if a vertex in V_2 is incident with the corresponding edge in $G(n,d)$. Notice that the graph is 2-regular on the vertices of V_1, and hence cannot have an expansion of greater than 1. This edge-vertex construction in [35] was applied to the Ramanujan graphs discussed below. It was noted by Alon that this edge-vertex construction is a special case of a construction be Ajtai et al [1] where vertices in one of the sets are identified with paths of length s and the vertices of the other with the vertices of the original graph. Edges in the new graph are then given by the path-vertex incidence on the original graph. The graphs are thus $(s+1)$-regular.

Thus good expanders (and magnifiers) are capable of leading to asymptotically good codes by a very simple construction. However the construction of such graphs is reminiscent of coding theory in the sense that a random construction of a graph with the required properties will almost surely lead to a good expander. Proving that such a randomly constructed graph is indeed a good expander and finding its expansion seems a difficult problem. However, one interesting approach to this problem is via the eigenvalues of the adjacency matrix of the graph.

The magnification or expansion of a graph is related to the eigenvalues of the adjacency matrix of the graph in the following manner. Let G be a k-regular connected graph of order n and adjacency matrix A_G. Then A_G has k as a simple eigenvalue and, as noted, if further G is bipartite, then $-k$ is also an eigenvalue and the eigenvalues are symmetric wrt the origin. In this case let $k > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n-1} \geq \lambda_{n-1}$. Let

$$\lambda = \lambda(G) = \max(|\lambda_i|)$$

where the maximum is taken over all eigenvalues $|\lambda_i| \neq k$. If $\lambda(G)$ is small then trivially so is λ_1 and, in particular, that $\lambda(G) = \lambda_1$ if G is bipartite. It can be shown that graphs with $\lambda(G)$ small are good expanders (and magnifiers). Thus one possibility to find a random graph and test whether or not it is a good expander by determining its eigenvalues.

The body of literature on graphs with small second eigenvalue is now extensive and a few of these papers are listed in the references, without citation, as of possible interest to the reader.

A basic and important result in this area is the following:

Theorem: If G_i *is a family of* k*-regular connected graphs with* k *fixed and order* $n_i = |G_i| \rightarrow \infty$ *then* $\liminf_{n_i \rightarrow \infty} \lambda(G_i) \geq 2\sqrt{k-1}$.

This result was first proved by Boppana and Alon although the first published proof appeared in [26]. Other proofs have since appeared (eg [32]).

A family of graphs that are asymptotically optimal with respect to the above criteria was found independently by Margulis [27] and Lubotsky et al [26], referred to as Ramanujan graphs. The name derives from a conjecture of Ramanujan on the number of ways to represent a number by a quadratic form in 4 variables. The graphs are Cayley relying on the quaternions and we note their properties. Let p and q be unequal primes congruent to 1 mod 4. The Ramanujan graphs $X^{p \cdot q}$ will be $p+1$-regular Cayley graphs formed from the group $PSL(2, \mathbb{Z}_q)$ if the Legendre symbol $\left(\frac{p}{q}\right) = 1$ and of $PGL(2, \mathbb{Z}_q)$ if $\left(\frac{p}{q}\right) = -1$. In both cases the Cayley graph is formed [26] from the $p+1$ generators chosen according to the $p+1$ ways of representing the prime p as a sum $p = {a_0}^2 + {a_1}^2 + {a_2}^2 + {a_3}^2$ with $a_0 > 0$. The construction obtained by Margulis [28] is closely related and perhaps more straight forward.

A brief hint as to the construction of these graphs is given, using the approach of [2]. Recall that $G = PSL(2, q)$ is the factor group of 2×2 matrices over $\mathbb{Z}_q$ with determinant 1 modulo the matrices $\{\pm I\}$. The graphs are the Cayley graphs of this group with respect to a subset of generators Ω defined below. Identifying vertices in the graph with the $q(q^2-1)/2$ elements of G, two such vertices $g,\ h$ will be adjacent if $gh^{-1} \in \Omega$. The set Ω is comprised of $p+1$ matrices formed in the following manner. To each of the $p+1$ ways it is possible express p as a sum of squares $a_0^2 + a_1^2 + a_2^2 + a_3^2 = p$ form the matrix:

$$g_{\underline{a}} = \pm \frac{1}{\sqrt{p}} \begin{pmatrix} a_0 + ia_1 & a_2 + ia_3 \\ -a_2 + ia_3 & a_0 - ia_1 \end{pmatrix}$$

where $i^2 \equiv -1 \pmod{q}$ and the square root exists. The set Ω is then comprised of the $p+1$ matrices so formed. The diffcult part of the construction is to prove the square root bound on the second eigenvalue which we omit.

In general a k-regular, connected graph G is said to be Ramanujan if $\lambda(G) \leq 2\sqrt{k-1}$. The Ramanujan graphs constructed are regular of degree $k = p+1$ with the following two possibilities:

Case 1: $\left(\frac{p}{q}\right) = -1$: $X^{p,q}$ is bipartite of order $n = |X^{p,q}| = q(q^2-1)$ and girth $g(X^{p,q} \geq 4\log_p q - \log_p 4$

Case 2: $\left(\frac{p}{q}\right) = 1$: $X^{p \cdot q}$ is not bipartite, of order $n = |X^{p \cdot q}| = q(q^2-1)$ and girth $g(X^{p,q}) = 2\log_p q$.

The result of Case 1 is particularly interesting: an infinite class of k-regular graphs which asymptotically satisfy $g \geq (4/3)\log_{k-1} n$ has been obtained. The construction of such graphs with large girth is an interesting problem in its own right, briefly discussed further in the next section. It is interesting to note that these graphs grew out of our interest in graphs with second eigenvalue as small as possible, to obtain as large an expansion as possible. The fact they also have large girth seems an added interest.

The work of Lubotsky et al [26] provided the lower bound on the girth

$$g(X^{p,q}) \geq 4\log_p(q) - \log_p(4).$$

An upper bound on the girth of Ramanujan graphs has also been given ([5]) where it is shown that

$$g(X^{p,q}) < 4 \log_p q + \log_p 4 + 2.$$

The lower bound of the girth in the construction of the Ramanujan graphs implies that $g = \gamma(\log_{k-1} n)$ and the constant γ is at most $4/3$.

In fact the exact girth of Ramanujan graphs can be determined [7, 29]. Notice that the interval of the lower and upper bounds on the girth of these graphs,

$$(4 \log_p q - \log_p 4, 4 \log_p q + \log_p 4 + 2)$$

is of length $2 + \log_p 16$ and since $p \geq 5$ and the girth for the bipartite case is even, the interval often contains only the one possibility. Indeed the exact expression for the girth of these graphs [7, 29] may be described as follows [23]. Define an integer to be *good* if it is *not* of the form $4^\alpha(8\beta + 7)$ for any nonnegative integers α, β (which from a theorem of Legendre means it is representable as a sum of three squares). Then

$$g(X^{p,q}) = \begin{cases} 2\lceil 2 \log_p q \rceil & \text{if } p^{\lceil 2 \log_p q \rceil} - q^2 \text{ is good} \\ 2\lceil 2 \log_p q + \log_p 2 \rceil & \text{otherwise} \end{cases}$$

The original construction of these graphs required p and q to be prime but both Alon et al [2] and Bien [4] note that it was assumed that q is a prime equal to 1 modulo 4 but this is unnecessary and the construction also works for q a prime power.

The literature on Ramanujan graphs is quite extensive and a few such papers are noted in the references, without citation in the paper, as of possible interest to the reader.

4 Graphs with Large Girth

The construction of infinite families of graphs with large girth with determined properties such as girth, has proven to be a challenging problem. Virtually all such constructions are either Cayley or geometric in nature. There are many interesting constructions but, apart from the remarkable Ramanujan graphs, there are few infinite families to draw on. A few examples of some graphs and families of graphs are noted here but the list is far from complete and is representative only.

Benson [3] used quadratic surfaces in four and six dimensional projective geometries over $\mathbb{F}_q$ to derive minimal regular graphs of degree $q + 1$ and girth 8 and 12 respectively.

Biggs and Hoare [8] gave a construction for cubic graphs based on a sextet construction and Weiss [39] later showed that these graphs have large girth with constant $4/3$.

Wenger [40] gave a geometric construction for certain extremal graphs that excluded cycles of certain lengths but only resolved the (small) girth in two cases.

Kou et al [17] derive large classes of LDPC matrices (and hence graphs) from point-line incidences in both projective and Euclidean geometries. This draws on earlier work in the sixties that show that many of these codes are in fact cyclic, hence attractive from an implementation point of view. It appears the girth of such graphs will be either 6 or 8.

Lazebnik et al [21] described the following class of graphs. Let $s \in (0, 1]$ such that q^s is an integer. Then a doubly infinite class of graphs is constructed such that $|V_1| = q^5$, $|V_2| = q^{3+s}$ with degrees q^s and q^2, respectively, and girth 8.

The doubly infinite sets of graphs $D(n, q)$ and their connected components $CD(n, q)$, constructed below, have the more interesting property that their girths increase with dimension [23]. To introduce the graphs $D(n, q)$ let P and L denote two copies of the countably infinite dimensional vector space over $\mathbb{F}_q$. Denote a point $(p) \in P$ and a line $[l] \in L$ by [20, 23]

$$(p) = (p_1, p_{11}, p_{12}, p_{21}, p_{22}, p'_{22}, p_{23}, \cdots, p_{ii}, p'_{ii}, p_{i,i+1}, p_{i+1,i}, p_{i+1,i+1}, p'_{1+1,i+1} \cdots)$$

and

$$[l] = (l_1, l_{11}, l_{12}, l_{21}, l_{22}, l'_{22}, l_{23}, \cdots, l_{ii}, l'_{ii}, l_{i,i+1}, l_{i+1,i}, l_{i+1,i+1}, l'_{1+1,i+1}. \cdots).$$

On the infinite bipartite graph $D(q)$ with $V_1 = P$ and $V_2 = L$ define the incidence relation by saying that point (p) is incident to the line $[l]$ if the following relationships hold (in $\mathbb{F}_q$):

$$\begin{aligned} l_{11} - p_{11} &= l_1 p_1 \\ l_{12} - p_{12} &= l_{11} p_1 \\ l_{21} - p_{21} &= l_1 p_{11} \\ l_{ii} - p_{ii} &= l_1 p_{i-1,i} \\ l'_{ii} - p'_{ii} &= l_{i,i-1} p_1 \\ l_{i,i+1} - p_{i,i+1} &= l_{ii} p_1 \\ l_{i+1,i} - p_{i+1.i} &= l_1 p'_{ii} \end{aligned}$$

where the last four relations are defined for all $i \geq 2$. For a positive integer n define P_n and L_n by projecting the spaces onto the first n coordinates. Adjacency between points and lines is then defined by the first $(n-1)$ relations given above (which involve only the first n coordinates) .

It was shown in [18] that these graphs $D(n,q)$ are disconnected for $n \geq 6$. The disconnected components are isomorphic and let one such connected component be designated $CD(n,q)$. Each such graph [24] is in fact q-regular, and for q odd has order exactly

$$2q^{n - \lfloor \frac{n+2}{4} \rfloor + 1}$$

and [12] girth $n+5$ for all odd n and all $q \geq 4$. There is also a technique to derive interesting induced subgraphs for these graphs that give flexibility for their use. Guinand and Lodge [15] have considered these graphs for LDPC's and examined their performance and the reader is referred there for further interesting comments on these graphs/codes.

5 Comments

A limited introduction to the application of some graph theoretic ideas to the construction of LDPC codes has been attempted in the hope that it might be of relevance to further work in the area. While the notion of expander graphs has met with success in the construction and performance of such codes, it is also probably true that any analysis depending on high girth may not be sufficient to show that the families of codes resulting are asymptotically good. On the other hand notions such as girth, expansion and regularity do play some role in the determination of code quality. In particular the recent and important contribution of Richardson et al [33] indicates that perhaps the deterministic construction of irregular bipartite graphs might be an important direction for further progress. Presumably the girths of the graphs constructed should not be too small.

References

[1] M. Ajtai, J. Komlós and E. Szemerédi, Deterministic simulation in LOGSPACE, 19th STOC, 132-140, 1987.

[2] N. Alon, J. Bruck, M. Naor and R. Roth, Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs, *IEEE Trans. Inform. Theory*, vol. 38, 509-516 (1992).

[3] C.T. Benson, Minimal regular graphs of girths eight and twelve, *Canad. J. math.*, vol. 26, 1091-1094 (1966).

[4] F. Bien, Constructions of telephone networks by group representations, *Notices of the Amer. Math. Soc.*, vol. 36, 5-22, 1989.

[5] N.L. Biggs, *Algebraic Graph Theory*, Cambridge University Press, Cambridge Tracts in Mathematics, Vol. 67, 1974.

[6] N.L. Biggs, Graphs with large girth, *Ars Combinatorica*, vol. 25C, 73-80 (1988).

[7] N.L. Biggs and A.G. Boshier, Note on the girth of Ramanujan graphs, *J. Comb. Theory, Ser. B,* vol. 49, 190-194 (1990).

[8] N.L. Biggs and M.J. Hoare, The sextet construction for graphs, *Combinatorica*, vol. 3, 153-165 (1983).

[9] P. Chiu, Cubic Ramanujan graphs, *Combinatorica*, vol. 12, 275-285 (1992).

[10] P. Elias, Error-free coding, *IRE Trans. Inform. Theory*, vol. PGIT-4, 29-37 (1954).

[11] J. Friedman, Some graphs with small second eigenvalue, *Combinatorica*, vol. 15, 31-42 (1995).

[12] Z. Füredi, F. Lazebnik, A. Seress, V.A. Ustimenko and A.J. Woldar, Graphs of prescribed girth and bi-degree, *Jour. Combin. Theory, Series B*, vol. 64, 228-239 (1995).

[13] R.G. Gallager, Low-density parity-check codes, *IEEE Trans. Inform. Theory*, vol. 8, 21-28 (1962).

[14] R.G. Gallager, *Low-Density Parity-Check Codes*, Cambridge, MA: MIT Press, 1963.

[15] P.S. Guinand and J. Lodge, Tanner type codes arising from large girth graphs, Queen's Symposium 1998.

[16] W. Imrich, Explicit construction of reguolar graphs without small graphs, *Combinatorica*, vol. 4, 53-59 (1984).

[17] Y. Kou, S. Lin and M. Fossorier, Low density parity check codes based on finite geometries: a rediscovery and more, preprint.

[18] F. Lazebnik, V.A. Ustimenko and A.J. Woldar, A new series of dense graphs of high girth, *Bull. Amer. Math. Soc,*, vol. 32, 73-79 (1995).

[19] F. Lazebnik and V.A. Lazebnik, New examples of graphs without small cycles and large size, *Europ. J. Combinatorics*, vol. 14, 445-460 (1993).

[20] F. Lazebnik and V.A. Ustimenko, Explicit construction of graphs with an arbitrary girth and of large size, *Disc. Appl. Math.*, vol. 60, 275-284 (1995).

[21] F. Lazebnik, V.A. Ustimenko and A.J. Woldar, New constructions of bipartite graphs on m, n vertices with many edges and without small cycles, *Jour. Combin. Theory, Series B*, vol. 61, 111-117 (1994).

[22] F. Lazebnik and P. Wang, On the structure of extremal graphs of high girth, *J. Graph Theory*, 147-153 (1997).

[23] F. Lazebnik, V.A. Ustimenko and A.J. Woldar, New upper bounds on the order of cages, *European J. Combinatorics*, vol. 4, (1997).

[24] F. Lazebnik, V.A. Ustimenko and A.J. Woldar, A characterization of the components of the graphs $D(k, q)$, *Discrete Mahtematics*, vol. 157, 271-283 (1996).

[25] F. Lazebnik, V.A. Ustimenko and A.J. Woldar, Properties of certain families of $2k$-cycle-free graphs, *J. Comb. Theory, Ser. B*, vol. 60, 293-298 (1994).

[26] A. Lubotzky, R. Phillips and P. Sarnak, Ramanujan graphs, *Combinatorica*, vol. 8, 261-277 (1988).

[27] G.A. Margulis, Explicit constructions of concentrators, *Problems of Information Transmission*, vol. 9, 325-332 (1972) (English Translation).

[28] G.A. Margulis, Explicit constructions of graphs without short cycles and low density codes, *Combinatorica*, vol. 2, 71-78 (1982).

[29] G.A. Margulis, Explicit group-theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators, *Problems of Information Transmission*, vol. 24, 39-46 (1988) (English Translation).

[30] M. Meringer, Fast generation of regular graphs and construction of cages, *J. Graph Theory*, vol. 30, 137-146 (1999).

[31] M. Morgenstern, Existence and explicit constructions of $q + 1$ regular Ramanujan graphs for every prime power q, *Jour. Comb. Theor, Series B*, vol. 62, 44-62 (1994).

[32] A. Nili, On the second eigenvalue of a graph, *Discrete Mathematics*, vol. 91, 207-210 (1991).

[33] T. Richardson, A. Shokrollahi and R. Urbanke, Design of good low-denstiy parity check codes, preprint, 1999.

[34] G. Royle, Cubic cages, `http://www.ss.uwa.edu.au/gordon/cages/index.html`.

[35] M. Sipser and D.A. Spielman, Expander codes, *IEEE Trans. Inform. Theory*, vol. 42, 1710-1722 (1996).

[36] R. M. Tanner, Explicit concentrators from generalized N-gons, *SIAM J. Alg. Disc. Math.*, vol. 5, 287-293 (1984).

[37] R.M. Tanner, A recursive approach to low-complexity codes, *IEEE Trans. Inform. Theory*, vol. 27, 533-547 (1981).

[38] R. M. Tanner, Minimum distance bounds by graph analysis, preprint.

[39] A. Weiss, Girths of bipartite sextet graphs, *Combinatorica*, vol. 4, 241-245 (1984).

[40] R. Wenger, Extremal graphs with no C^4's, C^6's or C^10's, *Jour. Comb. Theor, Series B*, vol. 52, 113-116 (1991).

[41] D.B. West, *Introduction to graph theory*, Upper Saddle River, NJ: Prentice-Hall Inc., 1996.

[42] P.-K. Wong, Cages - a survey, *J. Graph Theory*, vol. 6, 1-22 (1982).

Constructions of LDPC Codes using Ramanujan Graphs and Ideas from Margulis*

Joachim Rosenthal
Department of Mathematics
University of Notre Dame
Notre Dame, Indiana 46556-5683, USA
e-mail: Rosenthal.1@nd.edu

Pascal O. Vontobel
Signal and Information Processing Laboratory
ETH Zürich
CH-8092 Zürich, Switzerland
e-mail: vontobel@isi.ee.ethz.ch

October 3, 2000

Abstract

Some twenty years ago G.A. Margulis [8] proposed an algebraic construction of LDPC codes. In this paper we analyze the performance of the codes proposed by Margulis. Mimicking the construction of Margulis we describe a new powerful regular LDPC code whose construction is based on a Ramanujan graph.

1 Introduction

Low-Density Parity-Check (LDPC) codes were introduced by Gallager [3] and they have been the focus of intense research in recent years. Roughly speaking, an LDPC code is a binary linear block code having an $m \times n$ parity-check matrix H whose nonzero entries are sparse. LDPC codes are typically described by a bipartite graph. The n left-vertices $\{v_1, \ldots, v_n\}$ represent the code symbols and the m right-vertices $\{c_1, \ldots, c_m\}$ represent the code constraints. There is an edge between vertex v_j on the left and vertex c_i on the right whenever the entry $h_{i,j}$ of the matrix H is 1. With this the matrix H represents the adjacency matrix of the bipartite graph.

We say that an LDPC code is a *(λ, ρ)-regular code* if the degree d_v of every vertex v on the left is equal to the integer λ and the degree d_c of every vertex c on the right is equal to the integer ρ. For a (λ, ρ)-regular code the equality $n\lambda = m\rho$ must necessarily hold. In terms of the parity-check matrix H, regularity simply means that every row contains ρ entries of 1 and every column contains λ entries of 1.

Consider now a sequence of randomly chosen (λ, ρ) regular codes whose block length n is increasing. It has already been shown by Gallager [3, 12] that there exist decoding algorithms whose complexity remains linear in the block length. Moreover, in the limit the codes can be decoded up to a certain *threshold* near capacity.

---

*This research is supported in part by NSF Grant No. DMS-00-72383 and by grant TH-16./99-3.

Since the work of Gallager several authors attempted to construct explicitly LDPC codes. Probably the first significant work was done by Margulis [8]. In the last year several papers have appeared and we would like to mention [1, 5, 7, 16].

In order that an LDPC code is good several things are desirable:

1. It is desirable that the bipartite graph G has no small cycles. Recall that the girth of a graph is the length of its smallest cycle. For a given block length n and given degrees (λ, ρ) it is desirable that the girth is as large as possible. This will guarantee that e.g. the sum-product algorithm (see e.g. [4]) performs best. For bipartite graphs the girth is always an even number.
2. The graph should be a "good expander" [14]. This means that if one starts from any subset S of left-vertices whose size satisfies $|S| \leq \frac{m}{2}$ then the set $\partial(S)$ of vertices on the right connected to the vertices of S should satisfy:

$$|\partial(S)| \geq \varepsilon |S|$$

   where ε is a 'large expansion factor'.
3. It is desirable that the code has a good minimum distance. This is particularly important if the code is used at a high signal-to-noise ratio.
4. It is desirable that the code can be compactly described and that the encoding complexity is low.

The listed items are of course somehow conflicting. Item 4. can be achieved if the block length is chosen small but this will not result in a large distance nor in a large girth. Upper bounds for the minimum distance in terms of the rate and block length are well known. The following lemma provides an upper bound for the girth:

Lemma 1 *Consider a (λ, ρ)-regular LDPC code having block length $n = m\rho/\lambda$. Let $\alpha = (\lambda - 1)(\rho - 1)$. If the girth $c \equiv 2 \bmod 4$ then one has necessarily the inequality:*

$$c \leq 4 \log_\alpha m + 2 \tag{1}$$

Proof: We start at a right vertex and count the number of right vertices reached after at most $\frac{c-2}{4}$ steps. This results in the inequality:

$$1 + \rho(\lambda - 1) + \rho(\lambda - 1)\alpha + \cdots + \rho(\lambda - 1)\alpha^{\frac{c-2}{4}-1} \leq m.$$

From this we deduce the inequality

$$1 + \alpha + \cdots + \alpha^{\frac{c-2}{4}} = \frac{\alpha^{\frac{c-2}{4}+1} - 1}{\alpha - 1} \leq m.$$

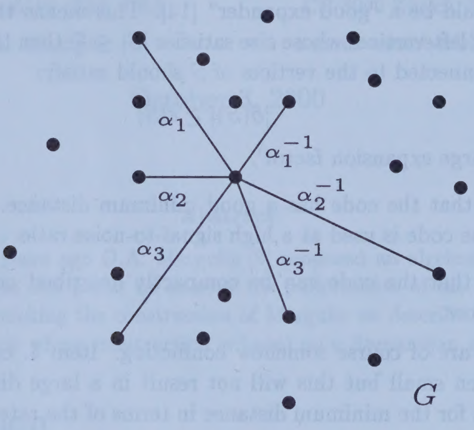
The result readily follows. □

In the next section we will study the performance of some LDPC codes introduced by Margulis [8]. In Section 3 we will use some known constructions of Ramanujan graphs [6, 9, 13] to design a remarkable $(3, 6)$-regular code, whose performance seems to be in a certain sense better than the performance of a randomly constructed code with the same design parameters. The code will have block length 4896 and almost optimal girth among all $(3, 6)$-regular codes. We have indications by simulations that the code has also a good minimum distance.

2 Cayley graphs and a construction of Margulis

One way of constructing k-regular graphs is by means of Cayley graphs. For this consider a finite group G. Let $\mathcal{A} \subset G$ be a subset of G satisfying $\mathcal{A} = \mathcal{A}^{-1}$. The Cayley graph $X(G, \mathcal{A})$ is the graph having as vertices the elements $g \in G$. The vertices $g, h \in G$ are connected by an edge whenever there is an $a \in \mathcal{A}$ such that $h = ga$. $X(G, \mathcal{A})$ is an undirected k-regular graph with $k = |\mathcal{A}|$.

Example 2 Assume the set $\mathcal{A}$ consists of $\mathcal{A} = \{\alpha_1, \alpha_1^{-1}, \alpha_2, \alpha_2^{-1}, \alpha_3, \alpha_3^{-1}\}$. Then one can think of the undirected graph as having the form:



Assume that the edges

$$(g_1, g_2), (g_2, g_3), \cdots, (g_{n-1}, g_n), (g_n, g_{n+1} = g_1)$$

form an n-cycle. By definition it follows that $g_1 = g_1 a_1 \cdots a_n$ or $a_1 \cdots a_n = e$, the identity element of the group G. This reasoning shows that a Cayley graph $X(G, \mathcal{A})$ has girth at least t if there are no non-trivial relations among the elements of $\mathcal{A}$ of length smaller than t.

One way of constructing a Cayley graph with good girth was given by Margulis [8]. For this let q be an odd prime. Let $\mathbb{F}_q$ be the finite field of q elements and consider the set $SL_2(\mathbb{F}_q)$ consisting of all 2×2 matrices $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with entries in $\mathbb{F}_q$ and having determinant $ad - bc = 1$. One readily verifies that $SL_2(\mathbb{F}_q)$ is a group of order $q^3 - q$. Consider the subset

$$\mathcal{A} := \left\{ A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, A^{-1} = \begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}, B^{-1} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \right\}.$$

Theorem 3 (Margulis [8]) *Let $G = SL_2(\mathbb{F}_q)$ and let $\mathcal{A}$ be as above. Then the Cayley graph $X(G, \mathcal{A})$ is a 4-regular graph with $q^3 - q$ vertices and girth*

$$c \geq 2 \log_\alpha (q/2) - 1, \quad \textit{where } \alpha = 1 + \sqrt{2} = 2.4142\ldots$$

In [8] Margulis showed how it is possible to construct from the graph $X(G, \mathcal{A})$ a $(3,6)$-regular LDPC code whose girth is at least as large as half the girth of the original graph. In the sequel we describe this construction and we investigate thereafter the performance of these codes for certain parameters.

As left vertices we will take two copies of G, say G and $\tilde{G}$. The right vertices of the bipartite graph will consist of the set G. An element $g \in G$ on the left will be connected with the right vertices

$$gA^2, gABA^{-1}, gB.$$

An element $\tilde{g} \in \tilde{G}$ on the left will be connected with the right vertices

$$\tilde{g}A^{-2}, \tilde{g}AB^{-1}A^{-1}, \tilde{g}B^{-1}.$$

If the smallest non-trivial relation among the elements $\{A, A^{-1}, B, B^{-1}\}$ has length c then one verifies that the elements

$$\{A^2, A^{-2}, ABA^{-1}, AB^{-1}A^{-1}, B, B^{-1}\}$$

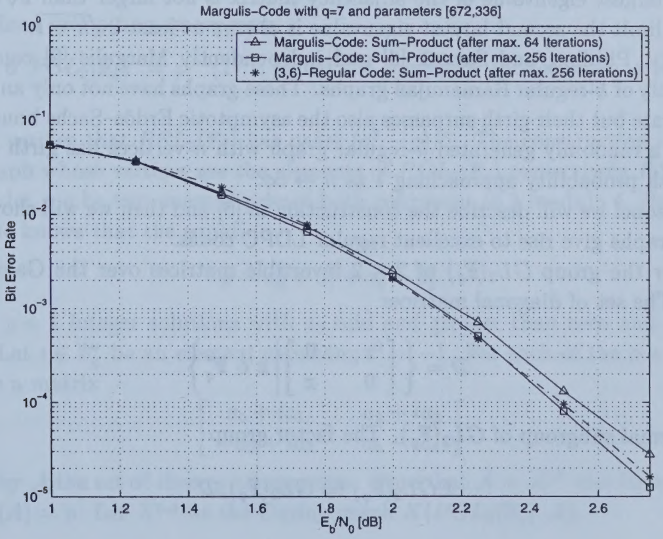
have no nontrivial relations of length smaller than $c/2 - 1$. It therefore follows:

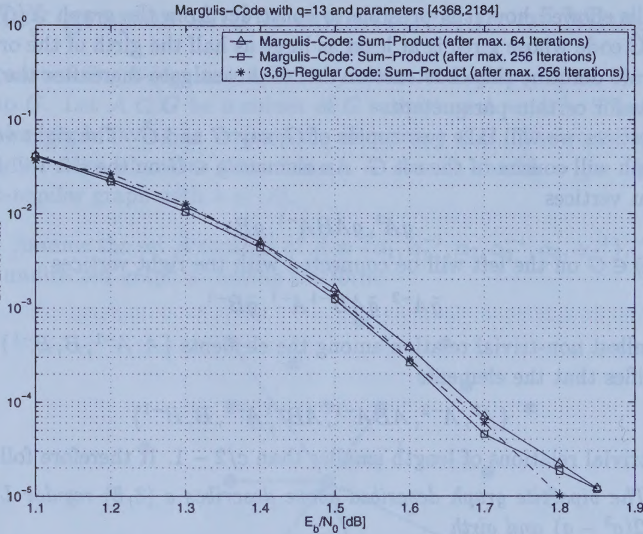
Lemma 4 *The bipartite graph described above describes a $(3,6)$-regular LDPC code of block length $2(q^3 - q)$ and girth*

$$c \geq \log_\alpha (q/2) - 1, \quad \textit{where } \alpha = 1 + \sqrt{2} = 2.4142\ldots$$

For small numbers of q the lower bound of the girth is of course not very impressive. Important is however the fact that the lower bound increases linearly in terms of $\log_\alpha n$, where n is the block length of the code. This is the best one can expect.

We found it interesting to simulate the performance of these codes for different parameters. We did simulations in the case when $q = 7$ and $q = 13$. When $q = 7$ we deal with a $(3,6)$ code of length $n = 672$ and dimension 336. When $q = 13$ we deal with a $(3,6)$ code of length $n = 4368$ and dimension 2184. We did perform the simulations using the sum-product algorithm (see e.g. [4])).





In the simulations the decoding process is terminated if the syndrome of the decoded codeword is zero or if the maximum number of iterations is reached.

3 A rate-1/2 code constructed from a Ramanujan graph

Ramanujan graphs are k-regular graphs which are in a certain sense optimal in their expansion behavior. Formally, a Ramanujan graph is defined through the property that the second largest eigenvalue of the adjacency matrix is not larger than $2\sqrt{k-1}$. Note that in the limit the second largest eigenvalue is always at least $2\sqrt{k-1}$.

Lubotzky, Phillips, and Sarnak [6] and independently Margulis [9] constructed an infinite family of k-regular Ramanujan graphs. These graphs have not only an exceptional expansion rate but their girth surpasses also the asymptotic Erdős–Sachs bound [2] which states that a randomly generated k-regular graph with n vertices has girth at least $c \geq \log_{k-1} n$ with probability approaching 1 as $n \to \infty$.

In the sequel we will describe the construction in [6] and then we will show how these k-regular graphs give rise to excellent regular LDPC codes.

Consider the group $GL_2(\mathbb{F}_q)$ of 2×2 invertible matrices over the Galois field of q elements. The set of diagonal matrices

$$\mathcal{D} = \left\{ \begin{bmatrix} x & 0 \\ 0 & x \end{bmatrix} \middle| \, x \in \mathbb{F}_q^* \right\}$$

forms a normal subgroup of $GL_2(\mathbb{F}_q)$. The factor group

$$PGL_2(\mathbb{F}_q) := GL_2(\mathbb{F}_q)/\mathcal{D}$$

is called the *projective general linear group*. Since $GL_2(\mathbb{F}_q)$ has $(q^2-1)(q^2-q)$ elements and since $\mathcal{D}$ has $q-1$ elements it follows that $PGL_2(\mathbb{F}_q)$ is a group of order q^3-q.

We can list the elements of $PGL_2(\mathbb{F}_q)$ in the following simple way:

1. There are $q^2(q-1)$ matrices having the form
$$\begin{bmatrix} 1 & b \\ c & d \end{bmatrix}, \text{ where } b, c \text{ arbitrary and } d \neq bc.$$
2. There are $q(q-1)$ matrices having the form
$$\begin{bmatrix} 0 & 1 \\ c & d \end{bmatrix}, \text{ where } d \text{ arbitrary and } c \neq 0.$$

These $(q^2+q)(q-1) = q^3-q$ matrices describe the group $PGL_2(\mathbb{F}_q)$ exactly.

For the rest of the paper we will assume that q is a prime. If $x \in \mathbb{F}_q^*$ is any nonzero element we define the Legendre symbol $\left(\frac{x}{q}\right)$ to be equal to 1 if x is a quadratic residue modulo q and as -1 if x is a quadratic non-residue.

Example 5 If $q = 17$ the quadratic residues of $\mathbb{F}_q^*$ are:

$$1, 4, 9, 16, 8, 2, 15, 13.$$

The quadratic non-residues are:

$$3, 5, 6, 7, 10, 11, 12, 14.$$

If $x, y \in \mathbb{F}_q^*$ then it is well known that $\left(\frac{x}{q}\right)\left(\frac{y}{q}\right) = \left(\frac{xy}{q}\right)$. As a consequence it follows that $\left(\frac{\det(xA)}{q}\right) = \left(\frac{x^2\det(A)}{q}\right) = \left(\frac{\det(A)}{q}\right)$ for every $A \in GL_2(\mathbb{F}_q)$ and $x \in \mathbb{F}_q^*$. It follows that

$$\varphi : PGL_2(\mathbb{F}_q) \longrightarrow \{-1, 1\}, \quad A \longmapsto \left(\frac{\det(A)}{q}\right)$$

is a well-defined group homomorphism.

Definition 6 $PSL_2(\mathbb{F}_q) := \varphi^{-1}(1)$ is called the *projective special linear group* over the field $\mathbb{F}_q$.

One readily verifies that $PSL_2(\mathbb{F}_q)$ has order $(q^3-q)/2$. The construction in [6] builds a Cayley graph whose vertices are the elements of $PGL_2(\mathbb{F}_q)$ respectively $PSL_2(\mathbb{F}_q)$.

For this, let p, q be two unequal primes both congruent to 1 modulo 4. By a theorem of Jacobi one knows that the equation

$$p = a_0^2 + a_1^2 + a_2^2 + a_3^2 \tag{2}$$

has exactly $p+1$ integer solutions with a_0 odd and greater than zero and a_j even for $j = 1, 2, 3$. Let $i \in \mathbb{F}_q^*$ be an element satisfying $i^2 = -1$. For each of the $p+1$ solutions of (2) define a matrix

$$\begin{bmatrix} a_0 + ia_1 & a_2 + ia_3 \\ -a_2 + ia_3 & a_0 - ia_1 \end{bmatrix}$$

and denote by $\mathcal{A}$ the set of these $p+1$ matrices. Note that $\mathcal{A} = \mathcal{A}^{-1}$ and for every $A \in \mathcal{A}$ one has $\det(A) = p$. Let $X^{p,q}$ be the Cayley graph $X(PGL_2(\mathbb{F}_q), \mathcal{A})$.

Theorem 7 ([6, 9]) *Let p, q be unequal primes congruent to 1 modulo 4 and satisfying $\left(\frac{p}{q}\right) = -1$. Note that $\left(\frac{p}{q}\right) = \left(\frac{q}{p}\right)$ in our situation. Then $X^{p,q}$ represents a bipartite graph having $q^3 - q$ vertices and girth at least*

$$c \geq 4 \log_p q - \log_p 4.$$

The graph $X^{p,q}$ is bipartite since we can take as left vertices the elements of $PSL_2(\mathbb{F}_q)$ and as right vertices the remaining group elements inside $PGL_2(\mathbb{F}_q)$. The homomorphism φ introduced above decides if an element belongs to left or to the right. Since the determinant of every element in $\mathcal{A}$ is a quadratic non-residue modulo q it follows that vertices on the left side are only connected to vertices on the right.

The asymptotic Erdős–Sachs bound [2] for a $(p+1)$-regular graph with $n = q^3 - q$ elements predicts that the girth of a randomly constructed graph will asymptotically satisfy $c \geq 3 \log_p q$. The girth of the graphs $X^{p,q}$ surpass the Erdős–Sachs bound for every pair p, q satisfying the assumption of the theorem. It was pointed out by Margulis [9] that the Erdős–Sachs bound has a flavor similar to the Gilbert–Varshamov bound as there exist also codes with a minimum distance which beats the average of randomly generated ones.

In the rest of the paper we show how the graphs $X^{p,q}$ give rise to interesting LDPC codes. We restrict ourselves to one example. During the preparation of this paper we learned that Lafferty and Rockmore [5] constructed also some LDPC codes starting from the Ramanujan graphs $X^{p,q}$. The construction approach taken by Lafferty and Rockmore differs however from the one presented next.

Let $p = 5$ and $q = 17$. Note that 5 is a quadratic non-residue in $\mathbb{F}_{17}$, i.e. $\left(\frac{5}{17}\right) = -1$. The graph $X^{p,q}$ consists in this case of $n = q^3 - q = 4896$ vertices.

Let $i := 4$ have the property that $i^2 = -1$ in $\mathbb{F}_{17}$. There are exactly $p+1 = 6$ solutions (a_0, a_1, a_2, a_3) having the property that

$$a_0^2 + a_1^2 + a_2^2 + a_3^2 = p = 5$$

with a_0 odd and positive and the others even, namely:

$$(1, \pm 2, 0, 0),\ (1, 0, \pm 2, 0),\ \text{and}\ (1, 0, 0, \pm 2).$$

This results in 6 matrices:

$$A^{\pm 1} := \begin{bmatrix} 1 \pm 8 & 0 \\ 0 & 1 \mp 8 \end{bmatrix},\ B^{\pm 1} := \begin{bmatrix} 1 & \pm 2 \\ \mp 2 & 1 \end{bmatrix},\ C^{\pm 1} := \begin{bmatrix} 1 & \pm 8 \\ \pm 8 & 1 \end{bmatrix}$$

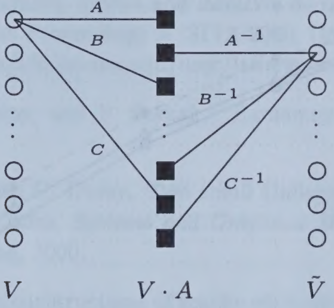
forming the set $\mathcal{A}$ of the Cayley graph. Note that the determinant is each time 5 modulo 17.

The graph $X^{p,q}$ has in a natural way a bipartite structure. The left node consist of $PSL_2(\mathbb{F}_{17})$, these are all elements whose determinant is a quadratic residue modulo 17. As right nodes take the elements whose determinant is a quadratic non-residue modulo 17. Note that a quadratic residue modulo 17 multiplied by 5 results in a quadratic non-residue modulo 17.

In order to build an LDPC code we take as left vertices two copies of $PSL_2(\mathbb{F}_{17})$, say V and $\tilde{V}$. As right vertices we take the right coset $VA \subset PGL_2(\mathbb{F}_{17})$ of $PSL_2(\mathbb{F}_{17})$.

As in the Margulis construction described in the last section we connect every element $v \in V$ on the left with the vertices vA, vB, vC on the right. An element $\tilde{v} \in \tilde{V}$ on the left will be connected with the vertices $\tilde{v}A^{-1}, \tilde{v}B^{-1}, \tilde{v}C^{-1}$ on the right.

The following diagram depicts the situation. The set VA describes the right coset of $PSL_2(\mathbb{F}_{17})$ in $PGL_2(\mathbb{F}_{17})$. The code bits are represented by the sets V and $\tilde{V}$. Note that for convenience the set $\tilde{V}$ has been drawn on the far right.

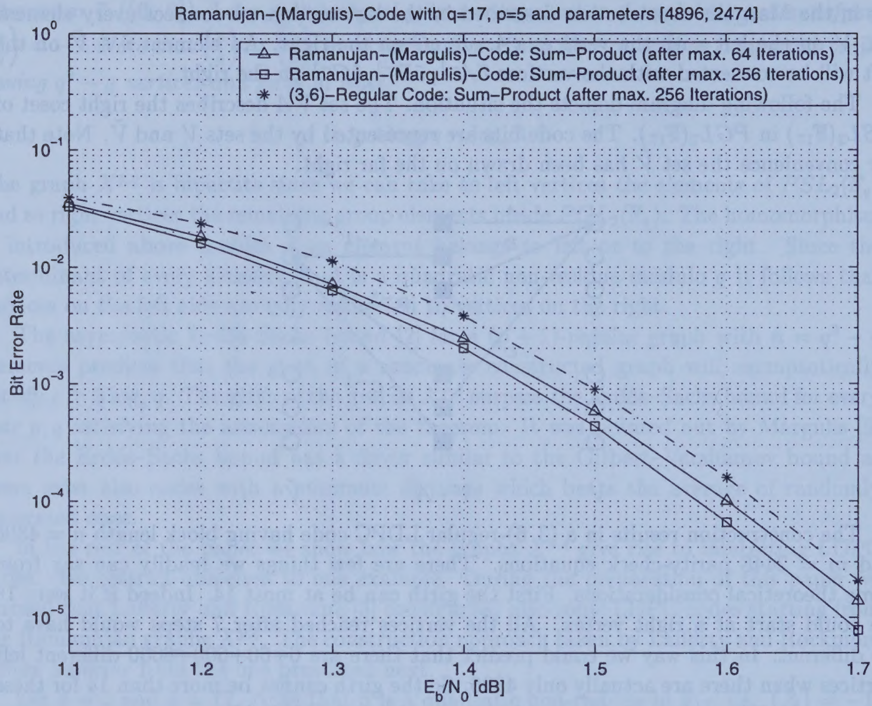


The construction results in a $(3,6)$-regular LDPC code having block length $n = 4896$ and $m = 2448$ parity-check equations. There are few things we readily can say from some theoretical considerations. First the girth can be at most 14. Indeed if it were 16 we could start at a right vertex. All the vertices reached after 7 steps would have to be different. In this way we could predict that there are 6+60+600+6000 different left vertices when there are actually only 4896. So the girth cannot be more than 14 for these parameters. We computed the actual girth as 12 using MATLAB.

Similarly one can make a counting argument for the minimum distance. An active parity check on the right is connected with two nonzero bits on the left. In two further steps these two bits are connected to 4 more nonzero bits. Completing this argument it follows that the minimum distance is at least 14. Actually we believe that the distance is much higher. Indeed during all our simulations, a large number of errors with large Hamming weight could be corrected.

Note that the resulting parity-check matrix has a significant number of dependent rows so that the dimension is 2474. (A difference of about 1% to the designed one). In our simulations we compare this with a randomly constructed $(3,6)$-regular code with parameters $[4896, 2448]$ which has the same decoding complexity per iteration.

The following diagram provides some simulation results.



The simulations suggest that the obtained algebraically constructed code performs superior over a randomly constructed (3, 6)-regular code.

4 Conclusion

In this paper we studied the performance of some algebraically constructed LDPC codes proposed by Margulis [8]. We adapted the construction using some Ramanujan graphs. It seems that these codes perform better than the randomly constructed codes of equal size and equal degrees.

References

[1] J. Bond, S. Hui, and H. Schmidt. Linear-congruence construction of low-density check codes. In B. Marcus and J. Rosenthal, editors, *Codes, Systems and Graphical Models*, IMA Vol. 123, pages 83–100. Springer-Verlag, 2000.

[2] P. Erdős and H. Sachs. Reguläre Graphen gegebener Taillenweite mit minimaler Knotenzahl. *Wiss. Z. Martin-Luther-Univ. Halle-Wittenberg Math.-Natur. Reihe*, 12:251–257, 1963.

[3] R.G. Gallager. *Low-Density Parity Check Codes.* M.I.T. Press, Cambridge, MA, 1963. Number 21 in Research monograph series.

[4] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. Submitted to IEEE Trans. Inform. Theory, available at http://www.comm.utoronto.ca/frank/factor/, 2000.

[5] J. Lafferty and D. Rockmore. Codes and iterative decoding on algebraic expander graphs. To appear in the Proceedings of ISITA 2000, Honolulu, Hawaii, available at http://www.cs.cmu.edu/afs/cs.cmu.edu/user/lafferty/www/, November 2000.

[6] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.

[7] D. J. C. MacKay and M. C. Davey. Two small Gallager codes. In B. Marcus and J. Rosenthal, editors, *Codes, Systems and Graphical Models*, IMA Vol. 123, pages 131–134. Springer-Verlag, 2000.

[8] G. A. Margulis. Explicit constructions of graphs without short cycles and low density codes. *Combinatorica*, 2(1):71–78, 1982.

[9] G. A. Margulis. Explicit group-theoretic constructions of combinatorial schemes and their applications in the construction of expanders and concentrators. *Problems Inform. Transmission*, 24(1):39–46, 1988. Translation from Problemy Peredachi Informatsii.

[10] M. Morgenstern. Existence and explicit constructions of $q+1$ regular Ramanujan graphs for every prime power q. *J. Combin. Theory Ser. B*, 62(1):44–62, 1994.

[11] T. Richardson, A. Shokrollahi, and R. Urbanke. Design of provably good low-density parity check codes. Submitted to IEEE Trans. Inform. Theory, 1999.

[12] T. Richardson and R. Urbanke. An introduction to the analysis of iterative coding systems. In B. Marcus and J. Rosenthal, editors, *Codes, Systems and Graphical Models*, IMA Vol. 123, pages 1–37. Springer-Verlag, 2000.

[13] P. Sarnak. *Some Applications of Modular Forms.* Cambridge University Press, Cambridge, 1990.

[14] M. Sipser and D. A. Spielman. Expander codes. *IEEE Trans. Inform. Theory*, 42(6, part 1):1710–1722, 1996.

[15] R. M. Tanner. A recursive approach to low complexity codes. *IEEE Trans. Inform. Theory*, 27(5):533–547, 1983.

[16] R.M. Tanner. Transforming quasi-cyclic codes with sparse graphs. Preprint, January 2000.

[17] J.-P. Tillich and G. Zémor. Optimal cycle codes constructed from Ramanujan graphs. *SIAM J. Discrete Math.*, 10(3):447–459, 1997.

Irregular Single Parity Check Product Codes

David Rankin
Dept. of Electrical and
Electronic Engineering
University of Canterbury
Private Bag 4800
Christchurch
New Zealand

T. Aaron Gulliver
Dept. of Electrical and
Computer Engineering
University of Victoria
P.O. Box 3055, STN CSC
Victoria, B.C.
Canada V8W 3P6

Abstract

Recently excellent performance has been achieved with irregular low density parity check (LDPC) codes [1, 2, 3, 4] and irregular turbo codes [5]. Here the performance of irregular single parity check (SPC) product codes and irregular randomly interleaved SPC product codes is investigated. The best degree profiles, for a specific rate loss and two degree sets, are determined by simulation for both classes of irregular product codes. Results over an ensemble of multiple degree sets show that performance is not very sensitive to the number of degrees.

1 Introduction

The aim of this work is to improve the performance of multi-dimensional single parity check (SPC) product codes [6] and randomly interleaved SPC product codes [6] by introducing irregularity into the structure of the code. This irregularity is best viewed by considering the graphical structure of the code.

In the sequel all SPC product codes have the same component code length, n, in each of the d dimensions [6]. The randomly interleaved SPC product codes [6] are SPC product codes in which the data (or data and parity) is interleaved between the encoding of each dimension. In both cases the product codes are sub-optimally decoded using iterative decoding of the component codes. Specifically the extrinsic information on each bit, in each dimension, is passed between the MAP decoders of the component codes [6].

The graphical structure of SPC product codes is such that every bit in the SPC product code is involved in exactly d SPC equations, and every equation contains exactly n bits. Thus from a decoding point of view, the variable and check nodes form a regular bipartite graph with degrees d and n, respectively. The variable nodes represent the bits which are transmitted (both data and parity) while the check nodes correspond to the SPC equations relating the appropriate bits. As an example, a (3,2) two-dimensional SPC product code is shown in Fig. 1.

A randomly interleaved SPC product code has a slightly different graphical structure due to the difference in encoding. The data bits still have degree d, where d is the number of dimensions. However the parity bits encoded in the q^{th} dimension have degree $d-q$ for $1 \leq q \leq d-1$, irrespective of the interleaving strategy (data only or data and parity bits). The degree of the parity bits, however, remains n for each SPC equation. Consequently a

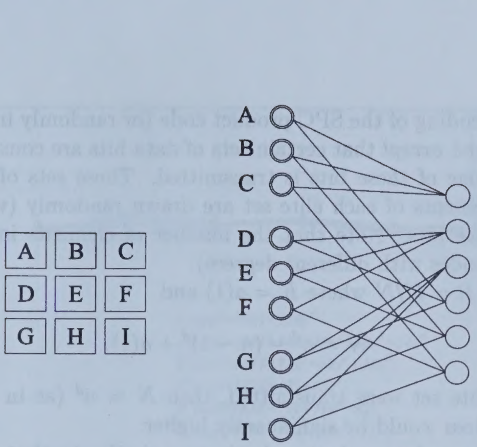


Figure 1: A two-dimensional (3,2) SPC product code (on the left), and the corresponding graphical structure on the right. The data bits are A,B,D,E and the others are parity bits.

randomly interleaved SPC product code is already a kind of irregular SPC product code (since the degree of the variable nodes is not constant). However the aim of this work is to improve the performance by adding more irregularity.

An irregular code will have a range of variable node degrees, and possibly also a range of check node degrees. Irregular LDPC codes are parameterized by the degree sequences (λ, ρ) [1] (related to the degrees of the variable and checks nodes), which are very useful when studying the asymptotic iterative decoding performance using *density evolution* [3, 4]. However, for our purposes it suffices to define the polynomial

$$a(x) = \sum_{i=1}^{d_v/d} a_i x^{i-1}, \tag{1}$$

where a_i is the number of variable nodes with degree id, and d_v is the maximum degree of any variable node. This polynomial defines the parameters of the code and can also be used to find the code rate (assuming that the check nodes remain unchanged).

2 Encoding and Decoding

It will be assumed that the length of the component codes, n, is constant, and hence any irregularity in the code will be added to the data bits within the variable nodes. This assumption allows the structure of the product code to remain unchanged, retaining some of the useful characteristics, and also reduces the number of variables which need to be optimised. To change the degree of the variable nodes, *data bits are repeated but only transmitted once.* Graphically this implies that the repeated data bits (corresponding to specific variable nodes), will form a "super node" since they represent the same bit. A super node is simply the union of the variable nodes corresponding to the data bits which have been repeated. Hence this new node has degree de where d is the number of dimensions and e is the number of times the bit is repeated throughout the data matrix. Note that this bit will only be transmitted once (so the rate loss is reduced), but in terms of decoding, the received bit is copied to the appropriate positions in the product code.

Therefore the encoding of the SPC product code (or randomly interleaved SPC product code) is unchanged *except* that certain sets of data bits are constrained to be exactly the same and only one of these bits is transmitted. These sets of data bits are called "elite" sets. The elements of each elite set are drawn randomly (without replacement) from all possible data bits. Note that the number of elements in each set may differ (resulting in super nodes with different degrees).

The code rate is $R = K/N$ where $K = a(1)$ and

$$N = n^d - (n-1)^d + a(1). \tag{2}$$

If all bits in each elite set were transmitted, then $N = n^d$ (as in the original product code), and the rate loss would be significantly higher.

Decoding these irregular product codes is very similar to that of the original codes [6], in which each dimension is MAP decoded in a particular order[1] and the extrinsic information on each bit is passed between the decoders. However there is one major difference, the elite sets (which correspond to super nodes in the graph), have much more extrinsic information available to them. Specifically in a normal product decoding algorithm the log likelihood ratio (LLR) of the *a priori* information, $L_q(X_k)$ (for the k^{th} bit in the q^{th} dimension) is defined to be the sum of the extrinsic information, $\mathbb{L}_{ki}$, in the other dimensions

$$L_q(X_k) = \sum_{\substack{i=1 \\ i \neq q}}^{d} \mathbb{L}_{ki}. \tag{3}$$

However in the case of a super node, the extrinsic information is also summed over all the other bits in the elite set. This assumes that the bits are independent which, due to the random choice of these bits, is generally a good assumption for the first few dimensions to be decoded. The effect of the extra information available for these elite bits is that they become very reliable, very quickly (with a high probability). This helps the other bits in SPC component codes which contain elite bits.

3 Results: Optimisation and Performance

The optimisation and performance results for both irregular SPC product codes and irregular randomly interleaved SPC product codes will be presented in separate sections. In both cases a significant improvement over the normal code can be achieved with a modest decrease in code rate. Typically the randomly interleaved codes have a better overall performance (for the same code rate and length), due to the reduced number of low weight codewords.

3.1 Irregular SPC Product Codes

This section begins with the performance (after 10 decoding iterations) of an ensemble of multiple degree four-dimensional (8,7) irregular SPC product codes at two different signal to noise ratios, $E_b/N_o = 1.29$dB and $E_b/N_o = 1.61$dB, which is shown in Fig. 2. The results in Fig. 2 represent irregular codes with different degree structures for each point on the x axis. For example, the results at $x = 1$ are for the structure, $a(x) = 1801 + 120x^4$,

---

[1] Decoding order is not important for a normal SPC product code but is for a randomly interleaved SPC product code.

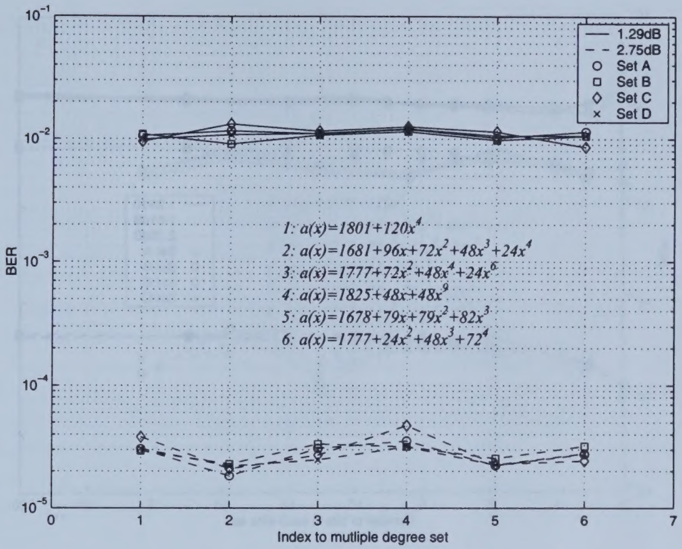


Figure 2: Simulation results for an ensemble of degree structures after 10 decoding iterations for the (8,7) four-dimensional irregular SPC product code.

and hence there are 120 elite sets each containing five data bits The sets A, B, C and D represent different choices of bits in each elite set (which are chosen at random without replacement). Note that the code rate and blocklength for each of the irregular codes in Fig. 2 are the same, specifically the rate is $R = 1921/(4096 - 2401 + 1921) = 0.531$ which is decreased from $(7/8)^4 = 0.586$. The degree structures were chosen to represent a wide variety of possible structures with the same overall code rate. The results obtained indicate that over the ensemble considered, most choices of irregular degree structures provide similar performance. We chose therefore to optimise the two degree irregular codes. Hence every data bit will be in either an elite set with e bits or it will not be repeated. This is clearly the simplest case to optimise since, for a given rate loss, we only need to change one parameter, e, which in turn will determine the number of elite sets containing these e data bits. The results of this optimisation is shown in Fig. 3. Clearly different choices of elite bits, for a given degree structure (sets A, B, C and D), have very similar performance. The optimum profile occurs when there are two bits in each of the elite sets, consequently $a(x) = 1441 + 480x$. This code has a large proportion of elite sets which will result in a significant reduction in the number of low weight codewords (one reason for the improved performance). The performance of the best two degree irregular SPC product code is compared with that for the normal SPC product code in Fig. 4. In both cases eight decoding iterations were performed. The irregular code shows a clear improvement. Note that the blocklength has decreased from 4096 to 3616 bits for the irregular code, but the code rate has also decreased from 0.586 to 0.531.

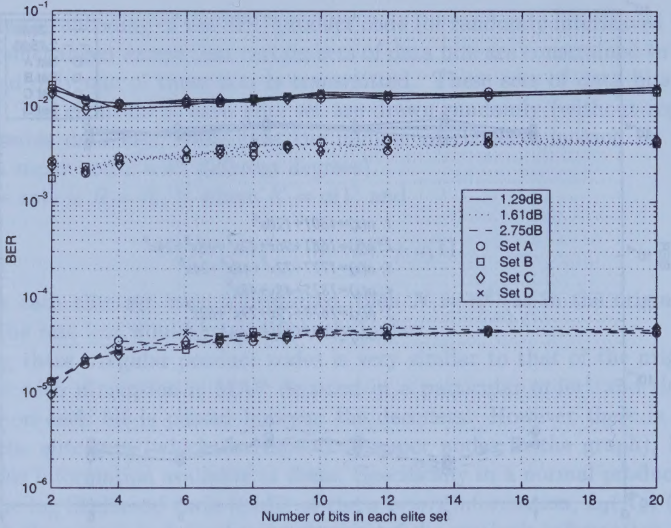


Figure 3: Simulation results for the two degree, (8,7) four-dimensional irregular SPC product codes after 10 decoding iterations.

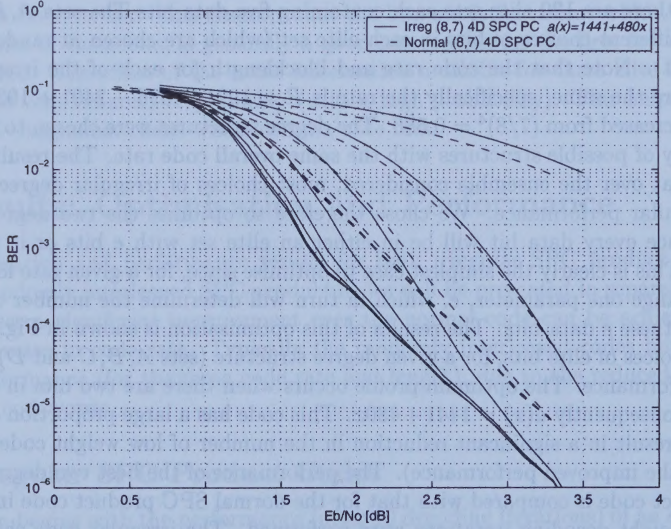


Figure 4: Performance comparison of the best two degree (8,7) four-dimensional irregular SPC product code ($a(x) = 1441 + 480x$), with a normal SPC product code. The curves represent 1 to 8 decoding iterations.

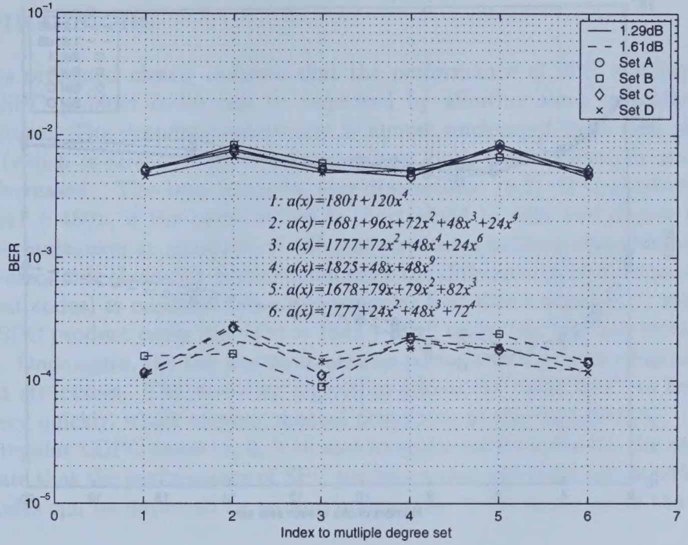


Figure 5: Simulation results for an ensemble of degree structures after 10 decoding iterations for the (8,7) four-dimensional irregular, randomly interleaved, SPC product code.

3.2 Irregular Randomly Interleaved SPC Product Codes

This section begins with the performance of an ensemble of multiple degree structures at 1.29dB and 1.61dB, which is shown in Fig. 5. The results in Fig. 5 show that most degree structures, on average, have similar performance. Note that the two degree structure performs among the best of those tested. This provides additional motivation for optimising the two degree class of irregular codes. Once again these two degree structures have only one variable to optimise, the number of bits in each elite set, e. The resulting optimisation for the two degree irregular codes, at $E_b/N_o = 1.29$dB and $E_b/N_o = 1.61$dB, is shown in Fig. 6. As before, the results corresponding to different choices of elite bits, sets A, B, C and D, are quite similar. Fig. 6 shows that six or seven is the optimum number of bits in each elite set for a code rate of 0.531. The optimum two degree (8,7) four-dimensional irregular randomly interleaved SPC product code has parameters $a(x) = 1841 + 80x^6$, and the performance of this code is shown in Fig. 7 along with the original (8,7) four-dimensional randomly interleaved SPC product code. It should be noted that unless all data bits are part of an elite set the minimum distance remains unchanged (2^d for the SPC product code and typically less for the randomly interleaved code). Consequently the reduction in the code rate, together with the same minimum distance, implies that the performance of the irregular codes will cross over that of the original code as the signal to noise ratio increases.

At least some of the improvement in performance can be attributed to the reduction in the number of low weight codewords. For example, in an SPC product code all weight one inputs result in a minimum distance codeword (but not all minimum distance codewords have input weight one), and only the fraction $a_1/a(1)$ of these input weight one minimum distance codewords will remain in the irregular code.

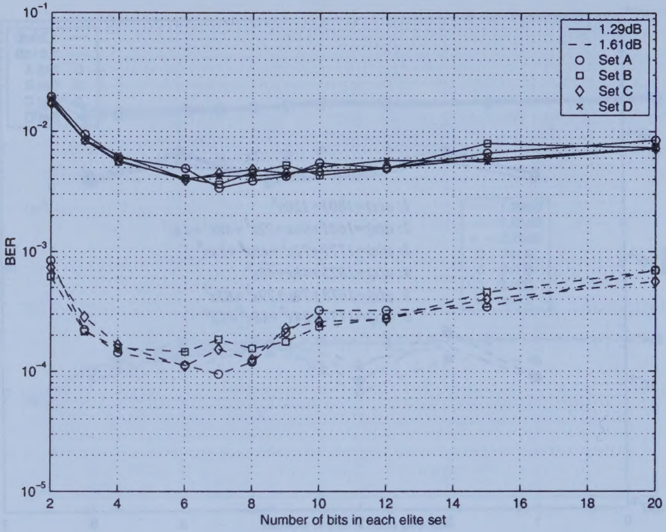


Figure 6: Simulation results for the two degree, (8,7) four-dimensional irregular, randomly interleaved, SPC product codes after 10 decoding iterations.

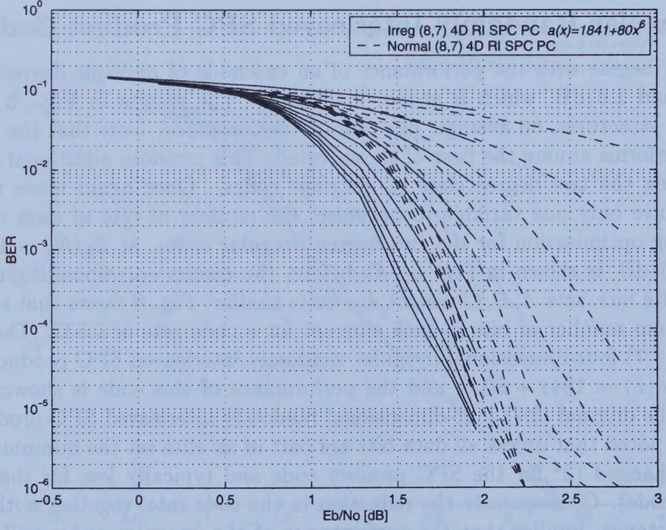


Figure 7: Performance comparison of the best two degree, (8,7) four-dimensional irregular, randomly interleaved, SPC product code ($a(x) = 1841 + 80x^6$), with a normal randomly interleaved, SPC product code. The curves represent 1 to 10 decoding iterations.

4 Conclusions

The results presented clearly indicate that the performance of SPC and randomly interleaved SPC product codes can be improved by allowing some irregularity in the code structure. The decoding complexity is almost unchanged from that of the original code (which is already very simple), however both the blocklength and the code rate are decreased. The best irregular four-dimensional (8,7) SPC product code has $a(x) = 1441 + 480x$, if the optimisation is constrained to only two degree structures. Simulation results over an ensemble of different structures indicate that this restriction is not an unreasonable choice for optimisation since the difference in performance (between the different codes) is minimal. The best irregular four-dimensional (8,7) randomly interleaved SPC product codes has $a(x) = 1841 + 80x^6$ under the two degree optimisation constraint. Once again, the two degree structures perform well compared to an ensemble of different structures. The decoding algorithm allows the "elite bits" to become very reliable, very quickly, which implies unequal protection of the data bits [8].

Like irregular LDPC codes [1, 2, 3, 4] and irregular turbo codes [5], the results given here indicate that the performance of SPC product codes and randomly interleaved SPC product codes can be improved by adding irregularity to the structure of the code.

References

[1] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi and D. A. Spielman, "Improved low-density parity-check codes using irregular graphs and belief propagation," Proc. *IEEE Int. Symp. Inform. Theory*, pp. 117, Aug. 1998.

[2] M. G. Luby, M. Mitzenmacher, M. A. Shokrollahi and D. A. Spielman, "Analysis of low density codes and improved designs using irregular graphs," Proc. *30th ACM STOC*, pp. 249–258, May 1998.

[3] T. Richardson, A. Shokrollahi and R. Urbanke, "Design of provably good low-density parity check codes," submitted to *IEEE Trans. Inform. Theory.*

[4] T. Richardson and R. Urbanke, "The capacity of low-density parity check codes under message passing," submitted to *IEEE Trans. Inform. Theory.*

[5] B.J. Frey and D.J.C. Mackay, "Irregular turbocodes," Proc. *37th Allerton Conf. on Commun. Control and Computing*, 1999.

[6] D. Rankin and T.A. Gulliver, "Single parity check product codes," submitted to *IEEE Trans. Commun.*, Dec. 1999.

[7] F.J MacWilliams and N.J.A. Sloane, *The Theory of Error Correcting Codes,* North-Holland, 1977.

[8] M.C. Davey, "Error-correction using low-density parity-check codes," PhD Thesis, University of Cambridge, Dec. 1999.

Performance Analysis and Code Optimization of Low Density Parity-Check Codes on Rayleigh Fading Channels*

Jilei Hou, Paul H. Siegel and Laurence B. Milstein
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093-0407
{jhou,psiegel,milstein}@cwc.ucsd.edu

Abstract

A numerical method has recently been presented to determine the noise thresholds of low density parity-check (LDPC) codes that employ the message passing decoding algorithm on the additive white Gaussian noise (AWGN) channel. In this paper, we extend this technique to the uncorrelated flat Rayleigh fading channel. Using a nonlinear code optimization technique, we optimized irregular LDPC codes for the uncorrelated Rayleigh fading channel. The thresholds of the optimized irregular LDPC codes are very close to the Shannon limit for this channel. For example, at rate 1/3, the optimized irregular LDPC code has a threshold only 0.08dB away from the channel capacity.

1 Introduction

Recent advances [1][2] in error correcting codes have shown that, using the message passing decoding algorithm, irregular low density parity-check (LDPC) codes can achieve reliable transmission at signal-to-noise ratios (SNR) extremely close to the Shannon limit on the additive white Gaussian noise (AWGN) channel, outperforming turbo codes of the same block size and code rate. LDPC codes have certain advantages, such as simple descriptions of their code structure and fully parallelizable decoding implementations. Moreover, they are amenable to theoretical decoding analysis. With iterative message passing decoders, LDPC codes exhibit an interesting threshold effect [1][3]: if the noise level of the channel is smaller than a certain threshold, the bit error probability goes to zero as the block size goes to infinity; if the noise level is above the threshold, the probability of error is always bounded away from zero. Richardson, *et al.* [1] developed a numerical technique, called *density evolution*, to analyze the performance of the message passing decoding algorithm on the AWGN channel, enabling the determination of thresholds to any desired degree of accuracy. In this paper, we extend this technique to the uncorrelated flat Rayleigh fading channel.

The code optimization of irregular LDPC codes is a nonlinear cost function minimization problem, a problem where differential evolution has been shown to be very effective and robust [4]. This technique has been successfully applied to the design of good irregular LDPC codes for both the erasure channel [4] and the AWGN channel [2]. We show that this technique is also very effective in the code optimization of irregular LDPC codes for the uncorrelated Rayleigh fading channel, and the threshold values of the optimized codes are extremely close to the capacity of this channel.

---

*This work was sponsored by the National Science Foundation under Grant NCR-9725568.

2 Low Density Parity-Check (LDPC) codes

An LDPC code is a linear block code specified by a very sparse parity-check matrix. As a linear block code, an LDPC code can be represented by a bipartite graph. Suppose the low density parity-check matrix H has N columns and M rows; the corresponding bipartite graph consists of N bit nodes, M check nodes, and a certain number of edges. Each bit node, called "left node," represents a bit of the codeword. Each check node, called "right node," represents a parity check of the code. An edge exists between a bit node and a check node if and only if there is a 1 in the corresponding entry in the parity-check matrix. We refer to the corresponding bit node and check node as the left and right neighbor nodes of the edge.

Regular LDPC codes are those for which all nodes of the same type have the same degree, where the degree of a node is the number of edges for which it is a neighbor node. A (j,k) regular LDPC code has a bipartite graph in which all bit nodes have degree j and all check nodes have degree k. Correspondingly, in the parity-check matrix H, all the column weights are j and all the row weights are k.

For irregular LDPC codes, the bit nodes (correspondingly the check nodes) can have different degrees. We say an edge has left (resp., right) degree i if its left (resp., right) neighbor node has degree i. An irregular LDPC code ensemble is specified by a degree sequence (λ, ρ) or its corresponding generating functions $\lambda(x) = \sum_{i=2}^{d_{lmax}} \lambda_i x^{i-1}$ and $\rho(x) = \sum_{i=2}^{d_{rmax}} \rho_i x^{i-1}$, where λ_i (resp., ρ_i) is the fraction of edges with left (resp., right) degree i and d_{lmax} (resp., d_{rmax}) is the maximal left (resp., right) degree of any edge.

3 Decoding Analysis

Firstly, we briefly review the decoding algorithm [2][5] for LDPC codes. For each edge of the underlying bipartite graph, the decoding algorithm iteratively updates two types of log a posteriori probability ratio (LAPPR) messages, q and r. The quantity q is the message sent from the bit node to the check node along a connecting edge e, which is expressed as $q = \log \frac{p(x=0|t)}{p(x=1|t)}$, where t denotes all the messages coming from the channel and the edges connected to the bit node, other than edge e. The quantity r is the message sent from the check node to the bit node along an edge e, which is defined as $r = \log \frac{p(x=0|v)}{p(x=1|v)}$, where v denotes the messages coming from the edges connected to the check node, other than edge e.

After l such iterations, the algorithm would produce the exact LAPPRs of all the bits if the bipartite graph defined by the parity-check matrix contains no loops of length up to $2l$ [5]. If we assume that the graph is loop-free, we can analyze the decoding algorithm directly because the incoming messages to every node are independent. Also, by the general concentration theorem of [1], for almost all the graphs in a code ensemble (λ, ρ) and almost all inputs, the decoder performance will converge to that of a corresponding loop-free graph as the codeword length approaches infinity.

Based on the assumption above, the following decoding analysis tries to track the *average* fraction of incorrect messages that is passed in each iteration of the decoding algorithm. Here, the fraction of incorrect messages is averaged over all the bits of a codeword.

First, we consider a regular (j, k) LDPC code. Since LDPC codes are linear block codes, we can assume that the all-zero codeword is sent and BPSK modulation $(0 \to 1,\ 1 \to -1)$ is used. Considering the message passed from the bit node to the check node, we have [2]

$$q = q_0 + \sum_{i=1}^{j-1} r_i\,, \tag{1}$$

where q_0 is the initial message conditioned on the channel output, and r_i, $i = 1, ..., j-1$, are the incoming LAPPR messages from all the incident edges, other than edge e. Since q_0 and r_i are independent random variables, the density function of q is the convolution of the density functions of all the elements in (1). This convolution can be efficiently computed in the Fourier domain. Let P_0 denote the density of q_0, P_l denote the density of q after l iterations, and R_l denote the density of r after l iterations. Letting F denote the Fourier transform operation,

$$P_l = F^{-1}\left(F(P_0)\left(F(R_{l-1})\right)^{j-1}\right), \tag{2}$$

where $R_0(r)$ can be set to Δ_0, and Δ_0 is defined as 1 if $r = 0$ and 0 if $r \neq 0$. We can write $P_l = P_l^1 + P_l^0$, where P_l^1 is supported on $(-\infty, 0]$ and P_l^0 is supported on $[0, \infty)$. Therefore, the fraction of incorrect messages after l iterations can be defined as

$$Pe(l) = \int_{-\infty}^{0} P_l(z)dz\,. \tag{3}$$

On the other hand, considering the message r passed from the check node to the bit node r, we have [2]

$$\tanh\frac{r}{2} = \prod_{i=1}^{k-1} \tanh\frac{q_i}{2}\,, \tag{4}$$

where q_i, $i = 1, ..., k-1$ are the incoming LAPPRs from the neighbor edges, other than edge e. To use the same method as described for the bit node to calculate the density function R_l, we need to apply logarithm operations on both sides of (4) to change the product into a summation,

$$\left(sgn_r\,,\ \log\left|\tanh\frac{r}{2}\right|\right) = \sum_{i=1}^{k-1}\left(sgn_{q_i}\,,\ \log\left|\tanh\frac{q_i}{2}\right|\right), \tag{5}$$

where the sign function $sgn_x = 0$ if $x \geq 0$, and $sgn_x = 1$ otherwise. Note that the summation for (5) is the mod-2 summation of the sign parts and the real summation of the magnitude parts. Therefore, the density of r can be computed in the Fourier domain in a manner similar to the computation of the density of q in (2).

This two-phase computation algorithm, called density evolution, makes it possible to track the fraction of incorrect messages, $Pe(l)$. At a certain noise level, we can run this algorithm iteratively until the error probability $Pe(l)$ either goes to zero or stops at a finite probability of error. The threshold, σ^*, denotes the supremum of all values of the noise level σ such that $\lim_{l\to\infty} Pe(l) = 0$, where σ is the standard deviation of the noise.

The density evolution algorithm can be extended to the irregular LDPC codes with only minor modifications, taking into consideration the irregular degree sequences; for example, at a bit node,

$$P_l = F^{-1}\left(F(P_0)\cdot\lambda\left(F(R_{l-1})\right)\right), \tag{6}$$

where $\lambda(x) = \sum_{i=2}^{d_{lmax}} \lambda_i x^{i-1}$ is as defined above.

For the uncorrelated Rayleigh fading channel, if the code symbol x is mapped into the signal point $w = (1-2x)$, the conditional pdf of the matched filter output y is

$$p(y|w,a) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-w\cdot a)^2}{2\sigma^2}\right), \tag{7}$$

where σ^2 is the variance of the noise, and a is the normalized Rayleigh fading factor with $E[a^2] = 1$ and density function $p(a) = 2a\exp(-a^2)$.

3.1 Ideal side information (SI)

When we have ideal SI, the initial message passed from the bit node to the check node is

$$q_0 = \log \frac{P(x=0|y,a)}{P(x=1|y,a)} = \frac{2}{\sigma^2} y \cdot a \,. \tag{8}$$

In the decoding analysis, assuming $w = 1$, q_0 has the conditional density function

$$P_0\left(q_0|a\right) = \frac{\sigma}{2a\sqrt{2\pi}} \exp\left(-\frac{\left(q_0 - \frac{2a^2}{\sigma^2}\right)^2}{8a^2/\sigma^2} \right) \,. \tag{9}$$

To get the unconditional density function of q_0, we average (9) over the density function of a,

$$\begin{aligned} P_0(q_0) &= \int_0^\infty \frac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\left(q_0 - \frac{2a^2}{\sigma^2}\right)^2}{8a^2/\sigma^2} \right) \exp(-a^2)da \\ &= \frac{\sigma}{\sqrt{2\pi}} \exp\left(\frac{-q_0(\sqrt{2\sigma^2+1}-1)}{2}\right) \int_0^\infty \exp\left(-\frac{\left(\frac{\sigma^2}{2a} q_0 - a\sqrt{2\sigma^2+1}\right)^2}{2\sigma^2} \right) da \,. \end{aligned} \tag{10}$$

3.2 No side information (no SI)

When no SI is available, following [6], we assume that $P(y|w)$ is Gaussian distributed in the region of the most probable y, and we approximate q_0 as

$$q_0 \approx \frac{2}{\sigma^2} y \cdot E[a] \,, \tag{11}$$

where $E[a] = 0.8862$. The corresponding conditional density function is

$$P_0\left(q_0|a\right) = \frac{\sigma}{2E[a]\sqrt{2\pi}} \exp\left(-\frac{\left(q_0 - \frac{2aE[a]}{\sigma^2}\right)^2}{8(E[a])^2/\sigma^2} \right) \,. \tag{12}$$

Averaging over the density function of a, we get

$$\begin{aligned} P_0(q_0) &= \frac{\sigma\Delta^2}{2E[a]} \exp\left(-\frac{\Delta^2\sigma^2 q_0^2}{4(E[a])^2}\right) \times \\ & \left[\sqrt{\tfrac{2}{\pi}} \exp\left(-\frac{\Delta^2 q_0^2}{8(E[a])^2}\right) + \frac{\Delta q_0}{E[a]} Q\left(\frac{-\Delta q_0}{2E[a]}\right)\right] \,, \end{aligned} \tag{13}$$

where $\Delta = \sqrt{\frac{\sigma^2}{2\sigma^2+1}}$ and $Q(x) = \frac{1}{2} erfc(\frac{x}{\sqrt{2}})$.

4 Consistency and Stability

Consistency is an important property associated with the message distribution in the density evolution. As defined in [2], a density function f on $[-\infty, \infty]$ is consistent if it satisfies $f(x) = f(-x)e^x$ for all $x \in [0, \infty]$, and it is shown that the initial message distributions for all the binary-input symmetric channels (including the AWGN channel) discussed in [2] satisfy this condition.

It was shown in [2] that the consistency property is invariant under density evolution, i.e., if P_0 is consistent, then the density functions of P_l and R_l calculated in density evolution are also consistent. It was next proved in [2] that if the density function of P_l is consistent, the average fraction of incorrect messages as defined in (3) is a non-increasing function of l and will always converge to a certain value, which might be zero.

In [2], the consistency property is then used to prove another important property of density evolution, which is summarized as follows: There exists an $\varepsilon > 0$ such that if density evolution is initialized with a consistent message density P_0 satisfying $\int_{-\infty}^{0} P_0(x)dx < \varepsilon$, the fraction of incorrect messages could converge to zero under density evolution if

$$\lambda'(0)\rho'(1) < e^{s} \tag{14}$$

where the parameter s is defined as

$$s = -\log(2\int_0^\infty P_0(x)e^{-x/2}dx)\,. \tag{15}$$

Conversely, if $\lambda'(0)\rho'(1) > e^{s}$, then the fraction of incorrect messages is strictly bounded away from 0.

In [2], (14) is referred to as the stability condition for the channel with initial message density P_0. For example, for the AWGN channel, the stability condition is given by [2]

$$\lambda'(0)\rho'(1) < e^{\frac{1}{2\sigma^2}} \tag{16}$$

We now show that the initial message density function of the uncorrelated Rayleigh fading channel with SI also satisfies the consistency property, and then we derive the stability condition for this channel. For the density function of the uncorrelated Rayleigh fading channel with SI, as defined in (10), it is easily verified that

$$\begin{aligned} P_0(q_0) &= \int_0^\infty \tfrac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\left(q_0-\frac{2a^2}{\sigma^2}\right)^2}{8a^2/\sigma^2}\right)\exp(-a^2)da \\ &= \exp(q_0)\int_0^\infty \tfrac{\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\left(-q_0-\frac{2a^2}{\sigma^2}\right)^2}{8a^2/\sigma^2}\right)\exp(-a^2)da \\ &= P_0(-q_0)\exp(q_0)\,, \end{aligned} \tag{17}$$

i.e., the initial message density function of the uncorrelated Rayleigh fading channel with SI satisfies the consistency condition. From (15) and (10), we have

$$\begin{aligned} e^{-s} &= 2\int_0^\infty P_0(q_0)e^{-q_0/2}dq_0 \\ &= 2\int_0^\infty da\tfrac{\sigma}{\sqrt{2\pi}}e^{-a^2}\int_0^\infty \exp\left(-\frac{\left(q_0-\frac{2a^2}{\sigma^2}\right)^2}{8a^2/\sigma^2}\right)e^{-q_0/2}dq_0\,, \\ &= \int_0^\infty 2a\exp(-(1+\tfrac{1}{2\sigma^2})a^2)da \\ &= 1/(1+\tfrac{1}{2\sigma^2}) \end{aligned} \tag{18}$$

i.e., the stability condition for the uncorrelated Rayleigh fading channel with SI is

$$\lambda'(0)\rho'(1) < 1 + \frac{1}{2\sigma^2}\,. \tag{19}$$

In the next section, we will numerically optimize the degree sequences for this channel, and we will verify empirically that they fulfill condition (19). As to the Rayleigh fading channel without SI, the initial density function (13) does not have the consistency property. We conjecture that this is because the expression (11) for the message, q_0, is only an approximation. Nevertheless, as shown in the numerical results, the density evolution technique for determining the thresholds still works quite well for this channel.

5 Results

5.1 Threshold calculation and code optimization

Using the density evolution technique discussed in Section 3, we can calculate the threshold values of LDPC codes for both the AWGN channel and the uncorrelated Rayleigh fading channel with or without SI. For convenience, in the following results, we will express the threshold by both σ and its corresponding (E_b/N_0)(dB). Since $\sigma^2 = \frac{1}{2R\cdot(E_b/N_0)}$, the threshold can also be defined as the smallest E_b/N_0 such that $\lim_{l\to\infty} Pe(l) = 0$. Fig. 1 compares the thresholds and the simulation results of rate-1/2 regular (3,6) LDPC codes on both the AWGN channel and the uncorrelated Rayleigh fading channel, where both ideal SI and no SI are considered. As shown, the thresholds for regular (3,6) LDPC codes on the AWGN channel, the fading channel with SI, and the fading channel with no SI, are 1.10dB, 3.06dB, and 4.06dB, respectively. The (3,6) LDPC codes used in the simulations are of length 10^5 and 10^6. The numerical threshold results are very consistent with the simulation results, and we conjecture that as the block size goes to infinity, the simulation results will converge to the thresholds. The results show that the regular (3,6) LDPC code suffers a loss of nearly 2dB and 3dB, respectively, in the fading channels with SI and without SI, relative to the AWGN channel.

Combining the density evolution and differential evolution techniques, we searched for good degree sequences with constraints on the maximal left degree d_{lmax} for the uncorrelated Rayleigh fading channel with SI. The differential evolution algorithm that we used in the code optimization is based on [4] with minor modifications. The resulting degree sequences of rate-1/3 codes for the Rayleigh fading channel with SI are shown in Table 1 for $d_{lmax} = 10, 16, 30, \text{and } 50$. Each column corresponds to one particular degree sequence. For each degree sequence, the coefficients of λ and ρ are given, as well as the threshold σ^*, and the corresponding $(E_b/N_0)^*$ in dB. Also listed is λ_2^*, the maximal value of λ_2 satisfying the stability condition (19). As can be seen, $\lambda_2 < \lambda_2^*$ for every degree sequence in the table, which confirms that these degree sequences satisfy the stability condition. We observe the higher the maximal left degree, the better the performance of the code.

Similar to the results that Richardson, *et al.* [2] obtained for the AWGN channel, the thresholds of the degree sequences optimized for the fading channel with SI are very close to the capacity of this channel (the capacity can be calculated following the method introduced in [7]). At rate 1/3, the capacity of the fading channel with SI is 0.4885dB. The degree sequence with $d_{lmax} = 50$ has the threshold of 0.5712dB, which is only 0.08dB away from the capacity!

For comparison, Table 2 gives the rate-1/3 code sequences that we optimized for the AWGN channel with $d_{lmax} = 10, 16, 30, \text{and } 50$. The threshold of the code sequence optimized for the AWGN channel with $d_{lmax} = 50$ is only 0.07dB away from the channel capacity.

It is interesting to see how the code sequences optimized for the fading channel with SI perform on the AWGN channel, and vice-versa. We compare the rate-1/3 code sequences optimized for the AWGN channel and the code sequences optimized for the fading channel with SI in Figs. 2. In Fig. 2(a), for each degree sequence, we show the gap between its threshold value for the AWGN channel and the AWGN channel capacity for rate 1/3. In Fig. 2(b), for each degree sequence, we show the gap between its threshold value for the fading channel with SI and the capacity for that channel for rate 1/3. It can be seen from Fig. 2(a) that the code sequences optimized for the fading channel with SI are also very good sequences on the AWGN channel, e.g., considering the degree sequence with $d_{lmax} = 50$ optimized for the fading channel with SI, its threshold value for the AWGN channel is only 0.16dB away from the channel capacity. Similar results can be observed from Fig. 2(b) for the reverse situation.

5.2 BER Simulation results

As shown in Fig. 1, the threshold values precisely predict the asymptotic performance as the block length of the LDPC codes approaches infinity. We are also interested in the performance of the optimized irregular LDPC codes when finite block size is considered. In the following results, we considered four rate-1/3 LDPC codes at a block size of 3072. The first one, LDPC ir1, is constructed according to the degree sequence with $d_{lmax} = 16$ optimized for the AWGN channel, as shown in Table 2. The degree sequence has thresholds of -0.32dB and 0.78dB on the AWGN channel and the fading channel with SI, respectively. Note that the capacities for these two channels are -0.4954dB and 0.4885dB, respectively. The second one, LDPC ir2, is constructed according to the degree sequence with $d_{lmax} = 16$ optimized for the fading channel with SI, as shown in Table 3, with thresholds of -0.25dB and 0.68dB on the AWGN and fading channel with SI, respectively. The third one, LDPC ir3, is constructed according to a degree sequence[1] with $d_{lmax} = 16$, which has thresholds of -0.18dB and 0.96dB on the AWGN and fading channel with SI, respectively. The last one, a quasi-regular[2] LDPC code R with all bit nodes degree-3, half of the check nodes degree-4, and half of the check nodes degree-5, has thresholds of 0.85dB and 2.13dB on the AWGN channel and the fading channel with SI, respectively. Among the four codes, ir1 has the best threshold on the AWGN channel and ir2 has the best threshold on the fading channel with SI.

Fig. 3(a) compares the simulation results of these four LDPC codes on both an AWGN channel and an uncorrelated Rayleigh fading channel with SI. It is shown that all three irregular LDPC codes achieve excellent performance on both channels. Note that the simulation results reflect the same relative performance as predicted by the computed threshold values. Furthermore, the irregular codes perform much better than the regular code. For example, at a BER of 2×10^{-4}, ir2 outperforms the regular LDPC code on the AWGN channel and fading channel with SI by about 0.7dB and 1.0dB, respectively.

In Fig. 3(b), ir1 and ir2 are compared to a (1,33/31,33/31) turbo code [8] of the same block size and the same code rate on both an AWGN channel and an uncorrelated fading channel with SI. We can see that ir1 and ir2 achieve virtually the same performance as the turbo code. On the AWGN channel, ir1 is slightly better than the other two codes, while on the fading channel with SI, ir2 is slightly better.

6 Conclusions

In this paper, we have shown that the numerical analysis technique for calculating the threshold of the LDPC codes for the AWGN channel can be extended to the uncorrelated flat Rayleigh fading channel. In addition, using the nonlinear optimization technique of differential evolution, we optimized the degree sequences for the uncorrelated Rayleigh fading channel and showed that the threshold values of these degree sequences are extremely close to the capacity of this channel. Simulation results for moderate block size showed that the optimized LDPC codes can achieve excellent performance on both the AWGN channel and the Rayleigh fading channel.

---

[1] This degree sequence was taken from Sae-Young Chung's web site: http://truth.mit.edu/~sychung/gaopt.html, which is an irregular LDPC codes design applet for design of good LDPC codes on AWGN channel.

[2] For rate-1/3 (j, k) regular LDPC codes, if we choose $j = 3$, we can not make k to be an integer. Therefore, we choose half of the check nodes degree-4 and half of the check nodes degree-5. We denote this code as a rate-1/3 *quasi-regular* LDPC code.

The authors wish to thank T. J. Richardson, R. Urbanke, S. Y. Chung, and H. D. Pfister for their helpful comments about the decoding analysis, and K. Tang for providing the simulation results for the turbo codes.

References

[1] T. J. Richardson and R. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," submitted to *IEEE Trans. Inform. Theory*.

[2] T. J. Richardson, A. Shokrollahi and R. Urbanke, "Design of provably good low-density parity-check codes," submitted to *IEEE Trans. Inform. Theory*.

[3] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA: MIT Press, 1963.

[4] A. Shokrollahi and R. Storn, "Design of efficient erasure codes with differential evolution," preprint, 1999.

[5] D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 399-431, March 1999.

[6] J. Hagenauer, "Viterbi decoding of convolutional codes for fading- and burst- channels," in *Proceedings of Int. Zurich Seminar on Digital Commun.'80*, pp. 1-7, Zurich, Switzerland, March 1980.

[7] E. K. Hall, S. G. Wilson, "Design and analysis of turbo codes on Rayleigh fading channels," *IEEE J-SAC*, vol. 16, no. 2, pp. 160-174, February 1998.

[8] K. Tang, P. H. Siegel and L. B. Milstein, "On the performance of turbo coding for the land mobile channel with delay constraints," in *Proceedings of 33rd Asilomar Conf. SSC'99*, pp. 1659-1665, Pacific Grove, CA, October 1999.

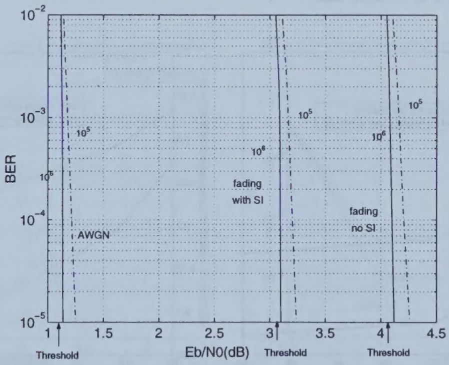
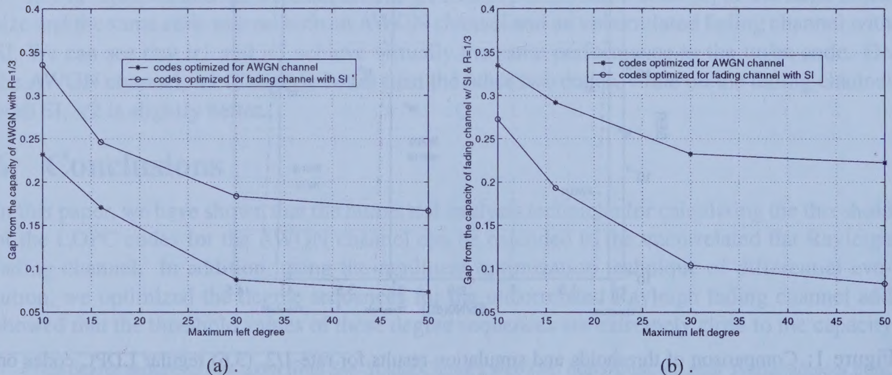


Figure 1: Comparison of thresholds and simulation results for rate-1/2, (3,6) regular LDPC codes on the AWGN channel, and the uncorrelated Rayleigh fading channels with or without SI. The codes used in simulations are of block size 10^5 and 10^6 .

d_{lmax}	10	16	30	50
λ_2^*	0.351321	0.312740	0.279123	0.248068
λ_2	0.328411	0.298433	0.267474	0.237738
λ_3	0.300171	0.245016	0.228605	0.200028
λ_4	0.001721			0.026700
λ_5		0.007173		
λ_6	0.001274	0.167659		
λ_7			0.140199	0.094505
λ_8			0.054426	0.050632
λ_9	0.000662		0.063206	0.028717
λ_{10}	0.367760		0.006163	0.060597
λ_{15}				0.057204
λ_{16}		0.281719		
λ_{29}			0.029347	
λ_{30}			0.210580	0.053621
λ_{49}				0.174248
λ_{50}				0.016010
ρ_4	0.024056	0.004024		
ρ_5	0.974961	0.547412	0.049352	
ρ_6	0.000983	0.448564	0.949897	0.538274
ρ_7			0.000751	0.359720
ρ_8				0.102006
$\sigma*$	1.1220	1.1323	1.1440	1.1468
$\left(\frac{E_b}{N_0}\right)^* dB$	0.7611	0.6817	0.5924	0.5712

Table 1: Good degree sequences of rate-1/3 for the uncorrelated Rayleigh fading channel with SI. For each sequence the threshold value $\sigma*$and the corresponding $\left(\frac{E_b}{N_0}\right)^*$ (dB) are given. The maximal value of λ_2 satisfying condition (19), λ_2^*, is given for $\sigma = \sigma^*$and the given $\rho'(1)$. Note that the capacity for this channel at code rate 1/3 is 0.4885 dB.



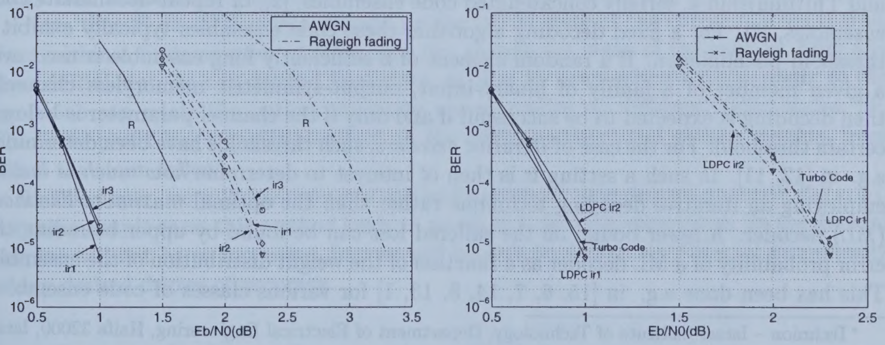
(a) .

(b) .

Figure 2: Comparison of the rate-1/3 code sequences optimized for the AWGN channel and the fading channel with SI. Note that the capacity for the AWGN channel at rate 1/3 is -0.4954dB and the capacity for the fading channel with SI at rate 1/3 is 0.4885dB.

d_{lmax}	10	16	30	50
λ_2^*	0.342056	0.298223	0.264470	0.237297
λ_2	0.329076	0.287567	0.256988	0.225792
λ_3	0.261590	0.230039	0.217847	0.207865
λ_4	0.048686	0.002147		0.012662
λ_5		0.068969		
λ_6		0.095590		
λ_7			0.163553	0.107496
λ_8		0.019108	0.061250	0.064003
λ_9			0.044084	0.032510
λ_{10}	0.360648		0.001780	0.012288
λ_{15}				0.100307
λ_{16}		0.296580		
λ_{29}			0.002474	
λ_{30}			0.252024	0.030314
λ_{50}				0.206763
ρ_4	0.000294	0.003426		
ρ_5	0.998683	0.424437		
ρ_6	0.001023	0.572137	0.882069	0.349540
ρ_7			0.114372	0.598609
ρ_8			0.003559	0.051851
$\sigma*$	1.2625	1.2714	1.2837	1.2858
$\left(\frac{E_b}{N_0}\right)^* dB$	-0.2637	-0.3247	-0.4084	-0.4225

Table 2: Good degree sequences of rate-1/3 for the AWGN channels. For each sequence the threshold value $\sigma*$ and the corresponding $\left(\frac{E_b}{N_0}\right)^*$ (dB) are given. The maximal value of λ_2 satisfying condition (16), λ_2^*, is given for $\sigma = \sigma^*$and the given $\rho'(1)$. Note that the capacity for this channel at code rate 1/3 is -0.4954 dB.



(a) .

(b) .

Figure 3: Simulation comparisons of rate-1/3, block size 3072, quasi-regular LDPC code R, irregular LDPC codes ir1, ir2, and ir3, and turbo code on both an AWGN channel and an uncorrelated Rayleigh fading channel with SI.

The asymptotic input-output weight distribution of convolutional encoders

Igal Sason* Emre Telatar† Rüdiger Urbanke‡

October 3, 2000

Abstract

We present a general method for computing the asymptotic input-output weight distribution of convolutional encoders. This method can be used to determine e.g. the (asymptotic) weight distribution of uniformly interleaved "turbo-like" code ensembles. In some instances, one can derive explicit analytic expressions. In general, to determine the growth rate for a particular normalized input weight κ and output weight ω, a system of polynomial equations has to be solved. We discuss several examples, give some general transformation rules, and discuss possible alternative approaches.

keywords: input-output weight distribution, maximum likelihood decoding, convolutional codes, "turbo-like" codes, uniform interleavers

1 Introduction

Consider the class of "turbo-like" code ensembles which are defined by convolutional encoders interconnected by *uniform* interleavers, [3]. Such ensembles include the standard parallel concatenated code ensembles [4] as originally introduced by Berrou, Glavieux and Thitimajshima, serially concatenated code ensembles, [2], or repeat-accumulate code ensembles, [7]. For a fixed decoding algorithm these code ensembles typically exhibit a threshold phenomenon. If a random element of a sufficiently long ensemble is used over a given member of a family of binary-input, output-symmetric memoryless channels, then decoding is expected to be successful if and only if the channel parameter is below a certain threshold. For the case of *iterative decoders*, such thresholds have been determined e.g. in [12, 11]. In such a setting it is then of interest to determine *how much is lost* by employing an iterative decoding technique rather than the optimal *maximum likelihood* (ML) decoder. A lower bound on the suffered loss can be found by upper bounding the error probability of a ML decoder as a function of the weight distribution of the ensemble. This has been done e.g. in [15, 6, 7, 14, 8, 13, 1] for various classes of code ensembles.

---

*Technion – Israel Institute of Technology, Department of Electrical Engineering, Haifa 32000, Israel , email:`eeigal@tiger.technion.ac.il`

†Swiss Federal Institute of Technology – Lausanne, LTHI-DSC, CH-1015 Lausanne, email:`Emre.Telatar@epfl.ch`

‡Swiss Federal Institute of Technology – Lausanne, LTHC-DSC, CH-1015 Lausanne, email:`Rudiger.Urbanke@epfl.ch`

A basic ingredient of this approach is to determine the average weight distribution of the given ensemble. Although there are analytic approaches to determine this weight distribution *exactly* for relatively small lengths [5, 10, 6], none of the suggested approaches allow for a direct determination of the asymptotic growth rate. The only exception is the case of repeat-accumulate codes, for which a compact analytic formula for the exact input-output weight distribution has been found in [7].

In this paper we will present a simple method that allows for the determination of the asymptotic input-output weight distribution of convolutional encoders. In a few simple cases we are able to derive analytic expressions. In general, the determination of the growth rate for a fixed normalized input weight κ and output weight ω requires the (numerical) solution of a system of polynomial equations.

2 Convolutional Encoders

Let F be a field and let $\mathsf{F}[[D]]$ denote the associated ring of rationals. By a *convolutional encoder* of *rate* $r := \frac{1}{t}$, $t \in \mathbb{N}$, we mean a t-tuple $G(D) := (G_1(D), \cdots, G_t(D))$, where $G_i(D) \in \mathsf{F}[[D]]$, $i \in [t]$.[1] All our examples will be limited to the case $\mathsf{F} = \mathsf{GF}(2)$ and $t \in \{1, 2\}$, although the basic method is not restricted to this setting. Let x be an input sequence of length l and let y be the corresponding output, where we restrict the output to the first l time steps, i.e., y will have total length lt. We will write $\mathrm{y} = \mathrm{x}G$ to denote this relationship. Consider the set

$$\mathcal{S}_G^l(\kappa, \omega) := \{(\mathrm{x}, \mathrm{y}) : \mathrm{y} = \mathrm{x}G; W(\mathrm{x}) = \kappa l; W(\mathrm{y}) = \omega t l\},$$

where $W(\mathrm{x})$ denotes the *Hamming weight* of x and κ, ω denote the *normalized* input and output weights, respectively. We are interested in the exponential growth rate of $|\mathcal{S}_G^l(\kappa, \omega)|$, which we denote by $F_G(\kappa, \omega)$. To make this growth rate well defined, consider the cardinality of the set of all input output pairs which have normalized weight within ϵ, $\epsilon > 0$, of κ and ω, respectively. One can check that this function has a well defined growth rate. Now define $F_G(\kappa, \omega)$ as the limit of this growth rate as ϵ approach zero. One can verify that this limit is well defined and that the limit function $F_G(\kappa, \omega)$ is continuous.

Note that $F_G(\kappa, \omega)$ does not depend on the initial state of the encoder. This is true, since within a finite number of steps, one can change the state of the encoder from any given initial state to any desired final state, so that by prepending a finite prefix to the input one can relate the output sequences corresponding to different initial states, [9]. The finite prefix has a negligible effect on the resulting input and output weights, from which the claim follows.

In this paper, we will give a general method to compute the desired quantity $F_G(\kappa, \omega)$. For some simple cases this method leads to an explicit analytic expression. In the general case we will not be so fortunate, but we will be able to determine $F_G(\kappa, \omega)$ numerically to any desired degree of accuracy.

3 General Transformation Rules

The following straightforward transformation rules allow in some cases to relate the exponential growth rates of two convolutional encoders. We say that a rational function

---

[1] The generalization of our results to encoders of rate $\frac{p}{q}$ is straightforward and we will omit the details.

$\frac{P(D)}{Q(D)}$ is in *minimal* form if $P(D)$ and $Q(D)$ are relatively prime. More generally, we will say that $G(D)$ is in *minimal* form if all its components $G_i(D)$ are in minimal form. Given a general encoder $G(D)$ we associate to it an encoder $\hat{G}(D)$ by eliminating from each component of $G(D)$ all common factors.

Lemma 1 [Transformation Rules]

[Minimal Form] $F_G(\kappa, \omega) = F_{\hat{G}}(\kappa, \omega)$.

[Shift] For any convolutional encoder $G(D) = (G_1(D), \cdots, G_t(D))$ of rate $\frac{1}{t}$ and any t-tuple of natural numbers $(s_1, \cdots, s_t)$,

$$F_{[D^{s_1}G_1(D), \cdots, D^{s_t}G_t(D)]}(\kappa, \omega) = F_{G(D)}(\kappa, \omega).$$

[Multiplicity] For any convolutional encoder $G(D)$ and any integer d,

$$F_{G(D^d)}(\kappa, \omega) = F_{G(D)}(\kappa, \omega).$$

[Duality] Let $G(D)$ be a convolutional encoder of rate one. Then

$$F_{\frac{1}{G}}(\kappa, \omega) = F_G(\omega, \kappa).$$

Proof. The minimal and the shift rule are elementary and are only listed for completeness. For the multiplicity rule, note that the output of the encoder $G(D^d)$ at the i-th time instance only depends on the input at time instances which congruent to i modulo d. Therefore, we can think of generating the output by first splitting the input into d separate streams and feeding each stream into an encoder $G(D)$ (rather than $G(D^d)$).[2] Finally consider the duality rule. Assume that $\mathrm{y} = \mathrm{x}G$ where $W(\mathrm{x}) = \alpha l$ and $W(\mathrm{y}) = \beta l$. Now feed y into the encoder $\frac{1}{G(D)}$. Then by definition, we will recover x (but perhaps for the last M steps, where M is the *memory length* of the encoder). This shows that $F_{\frac{1}{G}}(\beta, \alpha) \geq F_G(\alpha, \beta)$. The claim is now established by noting that by a reversal of the roles of $G(D)$ and $\frac{1}{G(D)}$, we get the converse inequality $F_G(\alpha, \beta) \geq F_{\frac{1}{G}}(\beta, \alpha)$. □

Example 2 $[G(D) = \frac{D}{1+D^2}]$ Consider the example $G(D) = \frac{D}{1+D^2}$. By the shift rule we have $F_{\frac{D}{1+D^2}}(\kappa, \omega) = F_{\frac{1}{1+D^2}}(\kappa, \omega)$, which by the multiplicity rule is equivalent to $F_{\frac{1}{1+D}}(\kappa, \omega)$. Using finally the duality rule, we conclude that $F_{\frac{D}{1+D^2}}(\kappa, \omega) = F_{1+D}(\omega, \kappa)$.
□

4 The Basic Method

We will now discuss the basic method of determining the quantity $F_G(\kappa, \omega)$. In some instances we will be able to derive analytic expressions for $F_G(\kappa, \omega)$. More generally though, the determination of $F_G(\kappa, \omega)$ for a fixed value of κ and ω will involve the solution of a system of polynomial equations which might have to be accomplished numerically.

Assume we are given a list $\mathcal{L} := \{(k_i, w_i, l_i)\}$ of triples. Each triple represents a *pattern* of *input weight* k_i, *output weight* w_i and *input length* l_i. Let lt be the given block length (length of the input stream).[3] We would like to determine in how many ways one can arrange these patterns (in a linear fashion) such that their total input weight is k their total output weight is w and their total input length is l. More precisely, we are interested in the *growth rate* of the quantity

---

[2] The problem that in the case of a $G(D^d)$ encoder the initial states are *linked* can be circumvented by inserting finite length guard sequences which guarantee that the encoder returns into the zero state.

$$\sum_{1\le m\le l}\ \sum_{\substack{\sum m_ik_i=k;\sum m_iw_i=w\\ \sum m_il_i=l;\sum m_i=m}} \binom{m}{m_1,\cdots,m_j}.$$

Define the *normalized* quantities $\kappa := \frac{k}{l}$, $\omega := \frac{w}{lt}$, $\mu := \frac{m}{lt}$ and $\mu_i = \frac{m_i}{lt}$. Note that we normalize the input weight by the the *input length* l, but the output weight as well as m and m_i by the *output length* lt. Then it is easily established that this growth rate is equal to

$$F_G(\kappa,\omega) := \max_{\substack{0\le\mu\le\frac{1}{t}\\ \sum\mu_ik_i=\kappa/t;\sum\mu_iw_i=\omega\\ \sum\mu_il_i=\frac{1}{t};\sum\mu_i=\mu}} \mu h(\frac{\mu_1}{\mu},\cdots,\frac{\mu_j}{\mu}),$$

where $h(\cdots)$ denotes the entropy function to the base of the natural logarithm.

Assume first that we fix μ. Using Lagrange multipliers we see that we have to find the stationary points of

$$\mu h(\frac{\mu_1}{\mu},\cdots,\frac{\mu_j}{\mu}) + \alpha\sum\mu_ik_i + \beta\sum\mu_iw_i + \gamma\sum\mu_il_i + \delta\sum\mu_i.$$

This suggests that the maximizing probability distribution has the form

$$\mu_i = \mu e^{\alpha k_i}e^{\beta w_i}e^{\gamma l_i}e^{\delta-1} \quad i=1,\cdots,j, \tag{1}$$

where the quantities α, β, γ and δ are chosen such that

$$\begin{align} \sum\mu_ik_i &= \frac{\kappa}{t}, \tag{2}\\ \sum\mu_iw_i &= \omega, \tag{3}\\ \sum\mu_il_i &= \frac{1}{t}, \tag{4}\\ \sum\mu_i &= \mu. \tag{5} \end{align}$$

Assume now that we specify the list $\mathcal{L}$ by means of the *formal power series*

$$R(X,Y,Z) := \sum_i X^{k_i}Y^{w_i}Z^{l_i},$$

where the triples (k_i, w_i, l_i) are not necessarily distinct. Then from (5) we get

$$\mu_i = \frac{\mu e^{\alpha k_i}e^{\beta w_i}e^{\gamma l_i}}{R(e^\alpha,e^\beta,e^\gamma)}. \tag{6}$$

Define

$$R_X(X,Y,Z) := X\frac{\partial R(X,Y,Z)}{\partial X} = \sum_i X^{k_i}Y^{w_i}Z^{l_i}k_i,$$

---

[3] The corresponding length of the output stream is then lt.

and in the same way define $R_Y(X,Y,Z)$ and $R_Z(X,Y,Z)$. Note that for μ_i of the form (6) we can express the left side of equation (2) as

$$\sum \mu_i k_i = \frac{\mu}{R(e^\alpha, e^\beta, e^\gamma)} \sum e^{\alpha k_i} e^{\beta w_i} e^{\gamma l_i} k_i = \mu \frac{R_X(e^\alpha, e^\beta, e^\gamma)}{R(e^\alpha, e^\beta, e^\gamma)}. \tag{7}$$

In a similar manner, we can express the left sides of the equations (3), (4) and (5) in terms of R_Y, R_Z and R. Therefore, if we let $x = e^\alpha$, $y = e^\beta$ and $z = e^\gamma$, then we can write equations (2)-(5) in the compact form

$$\begin{aligned} R_X(x,y,z) &= \frac{\kappa}{\mu t} R(x,y,z), & (8)\\ R_Y(x,y,z) &= \frac{\omega}{\mu} R(x,y,z), & (9)\\ R_Z(x,y,z) &= \frac{1}{\mu t} R(x,y,z). & (10) \end{aligned}$$

Note that every choice of (x,y,z) corresponds to a probability distribution on the list of patterns $\mathcal{L}$ and that each such choice has an associated entropy rate which, using equations (8-10), can be expressed as

$$\begin{aligned} E_R(x,y,z) &= -\mu \sum \frac{x^{k_i} y^{w_i} z^{l_i}}{R(x,y,z)} \ln\left(\frac{x^{k_i} y^{w_i} z^{l_i}}{R(x,y,z)}\right) \\ &= \mu \ln(R) - \frac{\kappa}{t} \ln(x) - \omega \ln(y) - \frac{1}{t} \ln(z) \\ &= \frac{R \ln(R) - R_X \ln(x) - R_Y \ln(y) - R_Z \ln(z)}{t R_z}. \end{aligned}$$

Summarizing the results so far we have

Theorem 3 Let $R(X,Y,Z) := \sum_i X^{k_i} Y^{w_i} Z^{l_i}$ be a formal power series specifying a set of patterns where the i-th pattern has associated input weight k_i, output weight w_i and length l_i. Let $\mathcal{S}_R^l(\kappa,\omega)$ denote the set of all linear arrangements of these patterns whose input weight is κl, whose output weight is $\omega t l$ and whose length is l. Let $F_R(\kappa,\omega)$ denote the growth rate of $|\mathcal{S}_R^l(\kappa,\omega)|$. Then

$$F_R(\kappa,\omega) = \sup_{\substack{(x,y,z) \in (\mathbb{R}^+)^3; \sum_i x^{k_i} y^{w_i} z^{l_i} < \infty \\ R_X(x,y,z)/R_Z(x,y,z) = \kappa \\ R_Y(x,y,z)/R_Z(x,y,z) = \omega t}} \frac{R \ln(R) - R_X \ln(x) - R_Y \ln(y) - R_Z \ln(z)}{t R_Z}.$$

If the formal power sum has a finite number of terms then the above formalism simply amounts to a compact description of the solution. The power of the approach appears when we allow formal power sums with an *infinite* number of terms which are *rational* generating functions! More precisely, we consider formal power series $R(X,Y,Z) = \sum_i X^{k_i} Y^{w_i} Z^{l_i}$ with non-negative integer coefficients such that $R(X,Y,Z) \in \mathbb{R}[[X,Y,Z]]$, the ring of rationals, or equivalently, $R(X,Y,Z) = \frac{P(X,Y,Z)}{Q(X,Y,Z)}$ for $P(X,Y,Z), Q(X,Y,Z) \in R[X,Y,Z]$, where $Q(0,0,0) \neq 0$ so that $Q(X,Y,Z)^{-1}$ exists.

Recall that we want to determine the growth rate of all input-output pairs which have normalized input weight κ and normalized output weight ω. To accomplish this, we need to find a generating function which counts *patterns* such that *every* codeword of the encoder can be parsed in a *unique* way into a sequence of these patterns and such that *any linear* arrangements of such patterns corresponds to a codeword. Let $\tilde{R}_G(X, Y, Z)$ denote the standard generating function of a convolutional encoder which counts *detours* from the zero state, i.e., codeword segments that start and end in the zero state and do not take on the zero state in between. Then a possible choice is $R_G(X, Y, Z) = Z + \tilde{R}_G(X, Y, Z)$ since every codeword is composed in a unique way of detours and "silent periods" in which the encoder rests in the zero state. But this choice is far from unique! Another possible choice is given by $R_G(X, Y, Z) = \frac{1}{1-Z}\tilde{R}_G(X, Y, Z)$. Here we think of patterns that consist of a "silent period" *together* with a detour. Although obviously both choices must lead to the same result, the involved algebra can be essentially different for the two cases. There are many more choices. We can count code segments that start and end in a given state $\mathbf{s}$ which is not necessarily the zero state. We will denote the corresponding generating function by $R_G^{[\mathbf{s}]}(X, Y, Z)$. These are only the most obvious choices, but many more degrees of freedom remain.

We are now ready to list our main result whose proof we omit for sake of brevity.

Theorem 4 For a given convolutional encoder $G(D)$ let $R_G^{[\mathbf{s}]}(X, Y, Z)$ have the form $R_G^{[\mathbf{s}]}(X, Y, Z) = \frac{P(X,Y,Z)}{Q(X,Y,Z)}$, where P and Q have no common factor and where X, Y and Z correspond to input weight, output weight and (input) length, respectively. Define

$$\mathcal{R}_R := \{(x, y, z) \in (\mathbb{R}^+)^3 : 0 < Q(x', y', z'),\ \forall (x', y', z'), 0 \le (x', y', z') \le (x, y, z)\},$$

and

$$\mathcal{K}_R(\kappa, \omega) := \{(x, y, z) \in \mathcal{R}_R : \frac{R_X(x, y, z)}{R_Z(x, y, z)} = \kappa;\ \frac{R_Y(x, y, z)}{R_Z(x, y, z)} = \omega t\}.$$

Then

(i) $\mathcal{R}_R$ is a convex set.

(ii) $F_G(\kappa, \omega) = \sup_{\mathcal{K}_R(\kappa,\omega)} E_{R_G^{[\mathbf{s}]}}(x, y, z)$.

We will now discuss some examples for which the above equations can be solved analytically.

Example 5 $[G(D) = 1 + D]$ We start with the simplest non-trivial encoder, namely $G(D) = 1 + D$. It is easily established that

$$R_{1+D}^{[0]}(X, Y, Z) = \frac{XY^2Z^2}{1 - XZ}.$$

A quick calculation shows that $R_Y = 2R$ and, therefore, from $R_Y = \frac{\omega}{\mu}R$ we conclude that $\mu = \frac{\omega}{2}$. Further, from $R_X = \frac{\kappa}{\mu}R = \frac{2\kappa}{\omega}R$ and $R_Z = \frac{1}{\mu}R = \frac{2}{\omega}R$, we get after some manipulations the solution $x^* = \frac{(2\kappa-\omega)(1-\kappa)}{\kappa(2-2\kappa-\omega)}$ and $z^* = \frac{2-2\kappa-\omega}{2-2\kappa}$. With these values we get

$$E_R(x^*, y, z^*) = (1 - \kappa)h\left(\frac{\omega}{2(1 - \kappa)}\right) + \kappa h\left(\frac{\omega}{2\kappa}\right),$$

which is independent of y. It follows that

$$F_{1+D}(\kappa, \omega) = (1 - \kappa)h\left(\frac{\omega}{2(1 - \kappa)}\right) + \kappa h\left(\frac{\omega}{2\kappa}\right).$$

Note that by our above remarks it follows that

$$F_{\frac{1}{1+D}}(\kappa,\omega) = (1-\omega)h\left(\frac{\kappa}{2(1-\omega)}\right) + \omega h\left(\frac{\kappa}{2\omega}\right).$$

As remarked in the introduction, the case $G(D) := \frac{1}{1+D}$ is actually the only non-trivial case in which $F_G(\kappa,\omega)$ was previously known. Even more, in this case it was shown in [7] that the *exact* number of input-output pairs of input weight k, output weight w and length l is equal to

$$\binom{l-w}{\lfloor k/2 \rfloor}\binom{w-1}{\lceil k/2-1 \rceil},$$

from which the asymptotic growth rate can be deduced directly. □

Example 6 $[G(D) = (1+D^2, 1+D+D^2)]$ Consider now the example $G(D) = (1+D^2, 1+D+D^2)$. After some manipulations we get

$$R^{[0,0]}_{(1+D^2,1+D+D^2)}(X,Y,Z) = \frac{XY^5Z^3}{(1-Z)(1-XYZ(1+Z))}. \tag{11}$$

A tedious but otherwise straightforward calculation reveals that

$$\begin{aligned}
\mu^* &= \frac{2\omega-\kappa}{8}, \\
z^* &= \begin{cases} \frac{\sqrt{5\kappa^2-10\kappa+8\kappa\omega+4(1-\omega)}-\kappa}{2(1-2\kappa)}, & \kappa < \frac{1}{2}, \\ \frac{\sqrt{5\kappa^2-10\kappa+8\kappa\omega+4(1-\omega)}+\kappa}{2(2\kappa-1)}, & \kappa > \frac{1}{2}, \end{cases} \\
x^*y^* &= \frac{1}{z^*(1+z^*)}\,\frac{5\kappa-2\omega}{4\kappa}.
\end{aligned}$$

Using these relations we can establish that

$$\begin{aligned}
F_{(1+D^2,1+D+D^2)}(\kappa,\omega) &= -\frac{2-\kappa-2\omega}{4}\ln(z^*) - \frac{5\kappa-2\omega}{8}\ln\left(\frac{5\kappa-2\omega}{4\kappa(1+z^*)}\right) - \\
&\quad \frac{2\omega-\kappa}{8}\ln\left(\frac{2\omega-\kappa}{4\kappa}(1-z^*)\right),
\end{aligned}$$

where the expression is valid in the region where $z^* < 1$ and $x^*y^* > 0$ and where the exponent is zero otherwise. A plot of $F_{(1+D^2,1+D+D^2)}(\kappa,\omega)$ is shown in Fig. 1. □

Example 7 $[G(D) = \frac{1+D}{1+D+D^2}]$ Consider now the example $G(D) = \frac{1+D}{1+D+D^2}$. In this case it is probably most convenient to look at

$$R^{[0,1]}_{\frac{1+D}{1+D+D^2}}(X,Y,Z) = Y^2Z^2\left(1+\frac{X^2Z}{1-Z}\right)\left(X+\frac{Z}{1-XZ}\right).$$

Note that $R_Y = 2R$ and therefore from the equation $R_Y = \frac{\omega}{\mu}R$ we get immediately $\mu = \frac{\omega}{2}$. Further, we get the following system of equations in x and z:

$$\begin{aligned}
\frac{2\,x^2\,z}{1+(-1+x^2)\,z} + \frac{x+\frac{x\,z^2}{(-1+x\,z)^2}}{x+\frac{z}{1-x\,z}} &= \frac{2\kappa}{\omega}, \\
z\left(\frac{1}{(1-x\,z)\,(x+(1-x^2)z)} + \frac{x^2}{(1-z)\,(1-(1-x^2)\,z)}\right) &= \frac{2}{\omega}-2.
\end{aligned}$$

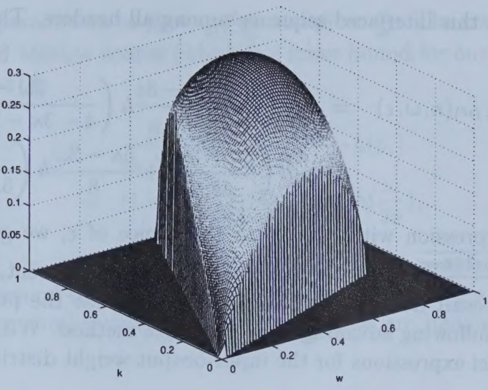


Figure 1: A plot of the growth rate $F_{(1+D^2,1+D+D^2)}(\kappa,\omega)$.

Note that $\mathcal{R}_R = \{(x,y,z) : 0 \leq z < 1; 0 \leq xz < 1\}$. For $\omega = \frac{1}{2}$ we get the analytic solution $x^* = \frac{\kappa}{1-\kappa}$, $z^* = 1-\kappa$ and then

$$F_{\frac{1+D}{1+D+D^2}}(\kappa, \frac{1}{2}) = h(\kappa).$$

For general ω it is not apparent how to solve this set of equations analytically. But it is fairly straightforward to solve the equations numerically and we get e.g. $F_{\frac{1+D}{1+D+D^2}}(0.2, 0.1) = 0.2959$, $F_{\frac{1+D}{1+D+D^2}}(0.01, 0.1) = 0.4917$. □

5 Alternative Approach

The basic method relies on finding generating functions for patterns such that every codeword can be parsed in a unique way into a sequence of these patterns, and such that any linear arrangement of these patterns results in a codeword. We now look at an alternative approach.

Example 8 [$G(D) = (1+D^2, 1+D+D^2)$ Revisited] Let's reconsider the case $G(D) = (1+D^2, 1+D+D^2)$. The combinatorial interpretation of the generating function given by (11) is the following. We have one *header* H which we can associate to XY^5Z^3 and three segments $\mathtt{T}_1 \leftrightarrow Z$, $\mathtt{T}_2 \leftrightarrow XYZ$, $\mathtt{T}_3 \leftrightarrow XYZ^2$. Further, $\mathtt{T}_2$ and $\mathtt{T}_3$ are *ordered*, i.e., they appear in an interleaved fashion. If we let μ, τ_1, τ_2, and τ_3 denote the normalized number of headers, and tail segments, respectively, then we get the relationships

$$\begin{pmatrix} 1 & 0 & 1 & 1 \\ 5 & 0 & 1 & 1 \\ 3 & 1 & 1 & 2 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix} = \begin{pmatrix} \frac{\kappa}{2} \\ \omega \\ \frac{1}{2} \end{pmatrix},$$

which has the general solution $\mu = \frac{2\omega-\kappa}{8}$, $\tau_1 = \frac{2-\kappa-2\omega}{4} - \epsilon$, $\tau_2 = \frac{5\kappa-2\omega}{8} - \epsilon$ and $\tau_3 = \epsilon$. How many ways are there now to arrange these patterns? Note that we can choose how to assign the segments $\mathtt{T}_1$ to the various headers, how to *interlace* the segments $\mathtt{T}_2$ and $\mathtt{T}_3$

and how to split up this interlaced sequence among all headers. This gives an exponent of

$$F_{(1+D^2,1+D+D^2)}(\kappa,\omega,\epsilon) \;:=\; \frac{4-3\kappa-2\omega-8\epsilon}{8}h\left(\frac{2\omega-\kappa}{4-3\kappa-2\omega-8\epsilon}\right)+ \frac{\kappa}{2}h\left(\frac{2\omega-\kappa}{4\kappa}\right)+\frac{5\kappa-2\omega}{8}h\left(\frac{8\epsilon}{5\kappa-2\omega}\right).$$

Maximizing this expression with respect to the choice of ϵ, we get the solution $\epsilon := \frac{2(1-\omega)-\sqrt{4(1-\omega)-10\kappa+5\kappa^2+8\kappa\omega}}{8}$. Although this is not readily apparent, this solution agrees with our previous result derived in Example 6. Note that the present method, when applicable, has the following advantage over our basic method. With slightly more effort we could derive *exact* expressions for the input-output weight distribution even for finite lengths! □

6 Lower Bounds

In some cases we might be content with finding non-trivial lower bounds. We can derive lower bounds based on our basic method if we are able to find at least one triple (x, y, z) which gives rise to the desired input and output weight. We will now discuss an alternative method of deriving lower bounds. The basic idea of the method is simple. Exhibit a source which, when connected to the encoder and in the limit of large blocklengths, gives rise to codewords with the desired normalized input and output weight. The entropy rate of the source is then a lower bound on the growth rate of such sequences. This method is best explained by discussing an explicit example.

Example 9 [$G(D) = 1 + D$ Revisited] Lets return to our favorite (since tractable) $1 + D$ encoder. Consider a *Markov source* which has two states as shown in Fig. 2. Hereby, the state signifies whether the last emitted bit was a zero or a one. We define here p_0 and p_1 to be the probabilities that the Markov source stays in the zero state or one state respectively (see Fig. 2) and we will use the shorthand $\bar{p_i} := 1 - p_i$, $i = 0, 1$. The steady state distribution of this Markov source is seen to be $(\frac{\bar{p_1}}{\bar{p_0}+\bar{p_1}}, \frac{\bar{p_0}}{\bar{p_0}+\bar{p_1}})$. Note that *essentially all* sequences which are emitted by this source have the following type: their normalized output weight is equal to $\frac{\bar{p_0}}{\bar{p_0}+\bar{p_1}}$ and their normalized number of consecutive pairs which differ is equal to $\frac{2\bar{p_0}\bar{p_1}}{\bar{p_0}+\bar{p_1}}$. Therefore, if we connect this source to the $1 + D$ encoder, then *essentially all* output sequences will have normalized output weight equal to $\frac{2\bar{p_0}\bar{p_1}}{\bar{p_0}+\bar{p_1}}$. Therefore, in order to get input-output pairs with normalized weights (κ, ω) we

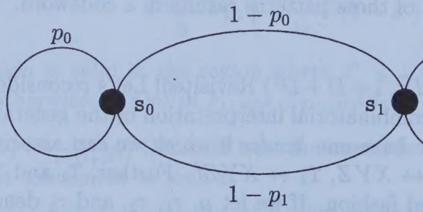


Figure 2: A two state Markov source.

need to choose the parameters as $\bar{p_0} = \frac{\omega}{2(1-\kappa)}$ and $\bar{p_1} = \frac{\omega}{2\kappa}$. The corresponding entropy rate of the associated Markov source (which is a lower bound for our desired growth rate by our remarks above) is

$$\begin{aligned} &\frac{\bar{p_1}}{\bar{p_0}+\bar{p_1}} h(\bar{p_0}) + \frac{\bar{p_0}}{\bar{p_0}+\bar{p_1}} h(\bar{p_1}) \\ &= (1-\kappa) h(\frac{\omega}{2(1-\kappa)}) + \kappa h(\frac{\omega}{2\kappa}), \end{aligned}$$

an old acquaintance. In this case we see that the proposed method leads to the exact result, although in general one can only claim that it will result in a lower bound on the growth rate. Many generalizations of the above method are imaginable and are worth exploring further. ☐

References

[1] S. Aji, H. Jin, A. Khandekar, R. J. McEliece, and D. J. C. MacKa, *BSC thresholds for code ensembles based on 'typical pairs' decoding.* Proceedings of IMA meeting. Available: `http://www.systems.caltech.edu/EE/faculty/rjm/`, 1999.

[2] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, *Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding'*, IEEE Trans. Inform. Theory, 44 (1998), pp. 909–926.

[3] S. Benedetto and G. Montorsi, *Unveiling turbo codes: Some results on parallel concatenated coding schemes*, IEEE Trans. Inform. Theory, 42 (1996), pp. 409–428.

[4] C. Berrou, A. Glavieux, and P. Thitimajshima, *Near Shannon limit error-correcting coding and decoding*, in Proceedings of ICC'93, Geneva, Switzerland, May 1993, pp. 1064–1070.

[5] F. Daneshgaran and M. Mondin, *An efficient algorithm for obtaining the distance spectrum of turbo codes*, in Proceedings of the International Symposium on Turbo Codes and Related Topics, Brest, France, Sept. 1997, pp. 251–254.

[6] D. Divsalar, *A simple tight bound on error probability of block codes with applications to turbo codes.* TMO Progress Report 42-139, JPL, pp. 1-35, Nov. 1999.

[7] D. Divsalar, H. Jin, and R. McEliece, *Coding theorems for "turbo-like codes"*, in Proceedings of the 1998 Allerton Conference, Monticello, Sept. 1998, pp. 201–210.

[8] T. M. Duman and M. Salehi, *New performance bounds for turbo codes*, IEEE Trans. on Communications, 46 (1998), pp. 717–723.

[9] J. Hokfelt, O. Edfors, and T. Maseng, *A survey on trellis termination alternatives for turbo codes*, in Proceedings 1999 IEEE 49th Vehicular Technology Conference, Houston, Texas, USA, May 1999, pp. 2225 – 2229.

[10] R. J. McEliece, *How to compute weight enumerators for convolutional codes*, in Communications and Coding, M. Darnel and B. Honary, eds., Tauton, Research Studies, 1998, ch. 6, pp. 121–141.

[11] T. RICHARDSON, A. SHOKROLLAHI, AND R. URBANKE, *Design of capacity-approaching low-density parity check codes.* accpeted for publication IEEE IT.

[12] T. RICHARDSON AND R. URBANKE, *The capacity of low-density parity check codes under message-passing decoding.* accepted for publication IEEE IT.

[13] I. SASON AND S. SHAMAI, *Improved upper bounds on the ML decoding error probability of parallel and serial concatenated turbo codes via their ensemble distance spectrum*, IEEE Trans. Inform. Theory, 46 (2000), pp. 24–47.

[14] S. SHAMAI AND I. SASON, *Variations on Gallager's bounding techniques: Performance bounds for turbo codes in gaussian and fading channels*, in Proceedings of the 2nd International Symposium on Turbo Codes and Related Topics, Brest, France, Sept. 2000, pp. 27–34.

[15] E. TELATAR AND R. URBANKE, *On the ensemble performance of turbo codes*, in International Symposium on Information Theory, Ulm, Germany, June 1997, IEEE, p. 105.

MAI Conditional Weak Convergence and Network Capacity in Wireless Networks with MMSE Receivers

Junshan Zhang
Electrical Engineering Dept.
Arizona State University
Tempe, AZ 85287-7206
junshan.zhang@asu.edu

Edwin K. P. Chong*
School of ECE
Purdue University
W. Lafayette, IN 47907
echong@ecn.purdue.edu

Abstract

We explore the performance of MMSE multiuser receivers in wireless networks where the signatures are modeled as random and take values in complex space. First we study the conditional distribution of the output multiple-access interference (MAI) of the MMSE receiver. By appealing to the notion of *conditional weak convergence*, we find that the conditional distribution of the output MAI, given the signatures and received powers, converges in probability to a proper complex Gaussian distribution that does not depend on the signatures. This result indicates that in a large network, the output interference of the MMSE receiver is approximately Gaussian with high probability, and that systems with MMSE receivers are robust to the randomness of the signatures. Building on the Gaussianity of the output interference, we then take the quality of service (QoS) requirements as meeting the signal-to-interference ratio (SIR) constraints and identify the network capacity of single-class systems with random spreading. The network capacity is expressed uniquely in terms of the SIR requirements and received power distributions. Compared to the network capacity corresponding to the optimal signature allocation, we conclude that at the cost of transmission power, the gap between the network capacity corresponding to optimal signatures and that corresponding to random signatures can be made arbitrarily small. Therefore, from the viewpoint of network capacity, systems with MMSE receivers are robust to the randomness of signatures.

1 Introduction

Consider a K-user communication system equipped with linear minimum-mean-square-error (MMSE) multiuser receivers. We focus on the following discrete-time synchronous baseband model [12, 13]: In a symbol interval, the received signal at the front end of the receiver is

$$Y^{(N)} = \sum_{i=1}^{K} \sqrt{P_i} b_i s_i + V, \tag{1}$$

---

*Supported in part by the National Science Foundation through grant ECS-9501652.

where the b_i's are the transmitted (complex) information symbols, the P_i's are the received powers, the s_i's ($s_i = \frac{1}{\sqrt{N}}[s_{i1}, \cdots, s_{iN}]^t$) are the signatures, and V comes from the sampling of the *proper complex* white Gaussian noise with power spectral density η.[1]

The signatures provide diversity gain and the corresponding model is of considerable interest (see, e.g., [12, 13]). In this paper, we assume that the signatures take values in complex space. Proper complex random processes arise naturally in equivalent baseband representations of bandpass communication systems [8]. In particular, for communications in fading environments it is sometimes more reasonable to assume for the baseband model that the signals are complex [13, Chapter 2]. We focus on systems with random spreading, as in [12, 14, 15, 16]. The random signature model is applicable to many practical systems. For example, this model is applicable to CDMA systems with very long pseudo-random spreading sequences, and also applicable to CDMA systems with random short signatures (i.e., the period of the signature equals the information symbol period and repetition of the same random signatures is adopted). In CDMA systems, the number of chips per signature N is sometimes called the *processing gain.* This model is also applicable to multiple-antenna systems where the vector s_i represents the fading levels of user i at each of the N antennas. In this case, we call s_i the *spatial fading signature* of user i. The s_i's in a multiple-antenna system are often modeled to be random and take values in N-dimensional complex space. The independence of the spatial fading signatures can be achieved by physically separating the antennas by a few carrier wavelengths [5]. Regardless of the specific applications, we call the length of the signatures the *degree of freedom.* More specifically, the degree of freedom is the processing gain in a CDMA system or the number of antennas in a multiple-antenna system [12].

Suppose there are K users in the system when the degree of freedom is N. We consider a more realistic scenario where users transmit data through a fading channel, and each user is capable of decentralized power control. Accordingly, we assume that the received powers are random (due to imperfect power control), and are independent across the users. We denote the received power of user i as P_i and its mean μ_i. Our results are asymptotic in nature, with both K and N going to infinity. As we scale up the system, the ratio of K to N is denoted by $\alpha \triangleq \frac{K}{N}$ and taken to be fixed, as is standard (see, e.g., [12, 14, 15]).

In this paper, we first study the distribution of the multiple-access interference (MAI) at the output of the MMSE receiver. The output MAI distribution is a physical layer performance metric and plays a crucial role in determining bit error probability. We focus on systems where repetition of the same random signatures is adopted and the received powers change relatively slowly compared to the symbol rate. In these systems it is of more interest to study the conditional distributions of the output MAI given the signatures and received powers. Our study makes use of the notion of *conditional weak convergence* [11]. In particular, our analysis involves two modes of conditional weak convergence—*convergence almost surely* of conditional distributions and *convergence in probability* of conditional distributions. Roughly speaking, our main result on the output MAI distributions can be summarized as follows:

- Assuming that the empirical distribution function[2] of $\{\mu_1, \cdots, \mu_K\}$ converges weakly as $N \to \infty$, the conditional distribution of the output MAI of the MMSE receiver, given the signatures and received powers, converges in probability to a proper complex Gaussian distribution that does not depend on the signatures.

---

[1] Proper complex random processes are also known as *circularly symmetric* random processes [3]. We elaborate further on proper complex random processes in Section 3.

[2] See [2, P. 279] and [1, P. 268] for the definitions of empirical distribution functions.

The above result can be viewed as a generalization of [16, 17], which established the Gaussianity of the output MAI under the assumption that the signatures are binary spreading sequences. However, the technical nature of the above result is in fact significantly different from that of [16, 17]. Indeed, the relaxation of the signatures to be complex gives rise to possibly much more variation in the MAI, requiring the use of the new notion of conditional weak convergence. We note that assuming signatures are deterministic, Poor and Verdú [9] have established the Gaussianity of the output interference of the MMSE receiver under several asymptotic conditions (the output MAI vanishes in these scenarios).

We then take a network perspective and identify the network capacity of single-class systems. Loosely speaking, a set of users is admissible if their simultaneous transmission does not result in violation of any of their quality of service (QoS) requirements; the *network capacity* is the maximum number of admissible users. Building on the Gaussianity of the output MAI, we take the QoS requirements as meeting the signal-to-interference ratio (SIR) constraints. Our result shows that the network capacity can be expressed uniquely in terms of the SIR requirements and received power distributions. The network capacity of imperfect power-controlled systems with linear receivers has been studied in [15]. In particular, in [15] the authors characterized the network capacity of systems with MMSE receivers for the deterministic signature case and the corresponding characterization for the random signature case was left open. We resolve this problem in this paper. Combining these results, we observe that at the cost of transmission power, we can drive the gap between the network capacity corresponding to optimal signatures and that corresponding to random signatures arbitrarily small.

The organization of the remainder of this paper is as follows. The next section contains our model description. In Section 3 we provide some necessary mathematical background. We present our main results in Section 4, and the corresponding proofs are omitted in this conference version. We draw our conclusions in Section 5.

2 Model Description

We focus primarily on a canonical discrete-time symbol-synchronous baseband model: In a symbol interval, the received signal before filtering is

$$Y^{(N)} = \sum_{i=1}^{K} \sqrt{P_i} b_i s_i + V.$$

Without loss of generality, we consider user 1. The MMSE receiver exploits the MAI structure provided by the signatures and received powers of the interferers. Because the received powers may vary from symbol to symbol, we assume that the MMSE receiver has knowledge of the mean received powers but not the instantaneous received powers of the interferers [16]. Also assume for now that the MMSE receiver has knowledge of P_1, the instantaneous received power of user 1. (It turns out that there is really no need for knowledge of P_1. We will return to elaborate on this issue in Section 4.) Define

$$\begin{aligned} S_1 &\triangleq [s_2, \cdots, s_K], & S &\triangleq [s_1, s_2, \cdots, s_K], \\ D_1 &\triangleq \operatorname{diag}(P_2, \cdots, P_K), & E_1 &\triangleq \operatorname{diag}(\mu_2, \cdots, \mu_K), \\ M_I &\triangleq S_1 E_1 S_1^H + \eta I, & M_I' &\triangleq S_1 D_1 S_1^H + \eta I. \end{aligned}$$

The MMSE receiver generates an output of the form $c^H Y^{(N)}$, where c is chosen to minimize the mean square error

$$J = \mathbb{E}\left[\,|c^H Y^{(N)} - b_1|^2 \big|\, P_1, S\right].$$

In what follows, we use the standard notation $\text{Re}\{z\}$, $\text{Im}\{z\}$, and $\bar{z}$ to denote the real part, imaginary part, and complex conjugate, respectively, of a complex number z. It is straightforward to see that the MMSE receiver is $\sqrt{P_1}M^{-1}s_1$, where $M \triangleq P_1 s_1 s_1^H + M_I$. After some algebra, the output at the MMSE multiuser receiver can be expressed as follows (cf. [4, 12, 15]):

$$y_1^{(N)} = \frac{P_1 s_1^H M_I^{-1} s_1}{1 + P_1 s_1^H M_I^{-1} s_1} b_1 + \mathcal{I}_1^{(N)} + \mathcal{I}_2^{(N)}, \tag{2}$$

where

$$\begin{aligned} \mathcal{I}_1^{(N)} &= \sum_{i=2}^{K} \frac{\sqrt{P_1}}{1 + P_1 s_1^H M_I^{-1} s_1} s_1^H M_I^{-1} \sqrt{P_i} b_i s_i, \\ \mathcal{I}_2^{(N)} &= \frac{\sqrt{P_1}}{1 + P_1 s_1^H M_I^{-1} s_1} s_1^H M_I^{-1} V. \end{aligned}$$

We note that $\mathcal{I}_1^{(N)}$ denotes the output MAI and $\mathcal{I}_2^{(N)}$ denotes the effect of background noise.

As is standard (see, e.g., [4]), the SIR is defined to be the ratio of the desired signal power to the sum of the noise and MAI powers at the output of the MMSE receiver in a symbol interval. Thus, we have that the SIR of user 1 is

$$\text{SIR}_1^{(N)} = \frac{P_1 (s_1^H M_I^{-1} s_1)^2}{s_1^H M_I^{-1} M_I' M_I^{-1} s_1}. \tag{3}$$

As will be shown in Theorem 4.1, in a large system the output interference of the MMSE receiver can be well approximated as Gaussian. Therefore, it is reasonable to take the QoS requirement as meeting the SIR constraints (see, e.g., [12, 14, 15]). Because of the randomness of the received powers and signatures, the SIR is random as well. Therefore, we adopt a probabilistic model for the users' QoS requirements as follows (cf. [7]):

$$P\{\text{SIR}_k^{(N)} \geq \gamma_k\} > a_k,$$

where $\text{SIR}_k^{(N)}$ is the SIR of user k when the degree of freedom is N, γ_k the *target SIR* of user k, and $a_k \in [0, 1)$, $k = 1, \cdots, K$. That is, the probability that the SIR of user k is no less than γ_k must be greater than a_k.

We are interested in identifying the maximum number of users admissible in a single-class system. A single-class system is one where the received powers of all the users are independently and identically distributed (see [12, 15]). Because the received powers are i.i.d. and the signatures are i.i.d., the SIR is identically distributed for all the users. Then it is clear that $P\{\text{SIR}_k^{(N)} \geq \gamma\}$ does not depend on k. Following [15], we define the *network capacity* of a system with linear MMSE receivers (when the degree of freedom is N) as follows:

$$\alpha_N(a) \triangleq \sup\left\{\alpha \,\middle|\, P\{\text{SIR}_1^{(N)} \geq \gamma\} > a\right\}.$$

The *asymptotic network capacity* $\alpha_\infty(a)$ is defined as the limit of $\{\alpha_N(a)\}$:

$$\alpha_\infty(a) \triangleq \lim_{N \to \infty} \alpha_N(a).$$

3 Mathematical Preliminaries

3.1 Two Modes of Conditional Weak Convergence

For convenience, we let $X^{(N)}$ denote the collection of the signatures and received powers when the degree of freedom is N, and $\boldsymbol{X}$ the sequence $\{X^{(N)}\}$. We are primarily interested in the asymptotic distributions of the $\mathcal{I}_1^{(N)}$'s given $\boldsymbol{X}$. Because the conditional distributions of the $\mathcal{I}_1^{(N)}$'s given $\boldsymbol{X}$ are random probability measures, the convergence of conditional distributions involves *conditional weak convergence.* In what follows, we refer to [11] and briefly restate a few definitions of modes of conditional weak convergence.

It is clear that the output MAI, $\mathcal{I}_1^{(N)}$, takes values in the complex space $\mathbb{C}$. Let M_1 be the space of all measures on $(\mathbb{C}, \mathcal{B}_\mathbb{C})$, where $\mathcal{B}_\mathbb{C}$ is the Borel σ-algebra on $\mathbb{C}$, and $\mathcal{M}_1$ is the σ-algebra generated by the weak topology on M_1. Let $C(\mathbb{C})$ denote the space of bounded continuous functions on $\mathbb{C}$. Let $P_N^{\boldsymbol{X}}$ denote the conditional distribution of $\mathcal{I}_1^{(N)}$ given $\boldsymbol{X}$. We define *almost sure convergence* of conditional distributions as follows:

Definition 3.1 *The conditional distribution of $\mathcal{I}_1^{(N)}$ given $\boldsymbol{X}$ converges almost surely to a random measure P_ν in $(M_1, \mathcal{M}_1)$, written as $P_N^{\boldsymbol{X}} \overset{a.s.}{\Longrightarrow} P_\nu$, if for all $f \in C(\mathbb{C})$, $\int f\, dP_N^{\boldsymbol{X}} \to \int f\, dP_\nu$ with probability one.*

As pointed out in [11], the most convenient formulation of *convergence in probability* of conditional distributions is in terms of almost sure convergence of subsequences of conditional distributions. We state its definition in the following.

Definition 3.2 *The conditional distribution of $\mathcal{I}_1^{(N)}$ given $\boldsymbol{X}$ converges in probability to a random measure P_ν in $(M_1, \mathcal{M}_1)$, written as $P_N^{\boldsymbol{X}} \overset{P}{\Longrightarrow} P_\nu$, if every subsequence $\{N'\}$ contains a further subsequence $\{N''\}$ for which $P_{N''}^{\boldsymbol{X}} \overset{a.s.}{\Longrightarrow} P_\nu$.*

Note that in the above, for brevity we have used the notation $\{N'\}$ and $\{N''\}$ to represent the subsequences $\{P_{N'}^{\boldsymbol{X}}\}$ and $\{P_{N''}^{\boldsymbol{X}}\}$, respectively.

Because $\mathbb{C}$ is separable and hence M_1 is metrizable, the above definition coincides with the standard definition of convergence in probability in a metric space. Indeed, this conclusion can be further illustrated by the following important result on convergence in probability of random variables $\{X_n\}$ [1, Theorem 20.5].

Lemma 3.1 *A necessary and sufficient condition for $X_n \xrightarrow{P} X$ is that each subsequence $\{X_{n'}\}$ contains a further subsequence $\{X_{n''}\}$ such that $X_{n''} \to X$ with probability one.*

3.2 Proper Complex Random Variables

We begin with a few definitions.

Definition 3.3 *A complex random variable is defined as a random variable of the form*

$$t = t_c + jt_s, \qquad j = \sqrt{-1},$$

where t_c and t_s are real random variables defined on the same probability space [6, p. 14].

As shown in [8], the "covariance" of two complex random variables $t = t_c + jt_s$ and $r = r_c + jr_s$ (where the "covariance" refers to the four covariances arising between the real and imaginary parts of t and r), when defined consistently with the corresponding notion for real random variables, is determined by the usual (complex) covariance together with a quantity called the *pseudo-covariance.* The covariance $\Sigma_{t,r}$ and pseudo-covariance $\tilde{\Sigma}_{t,r}$ are defined as follows:

$$\Sigma_{t,r} \triangleq \mathbb{E}[(t-\mathbb{E}[t])\overline{(r-\mathbb{E}[r])}]; \quad (4)$$

$$\tilde{\Sigma}_{t,r} \triangleq \mathbb{E}[(t-\mathbb{E}[t])(r-\mathbb{E}[r])]. \quad (5)$$

Definition 3.4 *A complex random variable $t = t_c + jt_s$ is* proper *if its pseudo-covariance $\tilde{\Sigma}_t = \mathbb{E}[(t-\mathbb{E}[t])^2]$ vanishes [8].*

Proper complex random variables are also known as *circularly symmetric* complex random variables. It should be noted that t is proper if and only if t_c and t_s have the same variance and are uncorrelated. As shown in [8], the complex multivariate Gaussian density assumes a natural form only for proper complex Gaussian random variables. We call the distribution of a proper complex Gaussian random variable a *proper complex Gaussian distribution.*

3.3 Regularity Conditions

We assume that the signatures are chosen randomly and independently. The model for random signatures is as follows: $s_i = \frac{1}{\sqrt{N}}[s_{i1}, \cdots, s_{iN}]^T$, where the s_{in}'s are i.i.d. proper complex random variables with zero mean and covariance 1, that is, $\mathbb{E}[|s_{in}|^2] = 1$. For technical reasons, we further assume that $\mathbb{E}[|s_{in}|^4] < \infty$.

We assume that the P_i's and the b_i's are independent. The assumptions we impose on the received powers and information symbols are listed as follows:

(3.P1) The empirical distribution function of $\{\mu_1, \cdots, \mu_K\}$ converges weakly to a distribution function H_μ as $N \to \infty$;

(3.P2) The P_i's are uniformly bounded above and the μ_i's are bounded below by a positive number;

(3.P3) The b_i's are independent proper complex random variables with $\mathbb{E}[b_i] = 0$ and $\mathbb{E}[|b_i|^2] = 1$, and the fourth moments of the b_i's are bounded.

We remark that Condition (3.P3) on the signal constellation covers many modulation methods of interest, for example, QPSK modulation schemes. It is clear that the quantity V in (1) is a proper complex random vector because any vector of samples taken from a proper complex random process is also proper [8].

4 Summary of Main Results

In this section, we summarize the main contributions of this paper.

Our first main result is on the asymptotic distributions of the output MAI conditioned on the signatures and received powers. This result is a generalization of [17], which established the Gaussianity of the output MAI under the assumption that the signatures are binary spreading sequences. We note that although the proof of part (a) has some similarity in flavor to that of [17, Theorem 3.1], the technical nature of our proof for part (b), the main part of Theorem 4.1, is in fact significantly different from that of [17, Theorem 3.2]. Indeed, the relaxation of the signatures to be complex gives rise to possibly much more variation in the MAI. We use conceptually the more subtle notion of *conditonal weak convergence* to reslove this problem.

Theorem 4.1 *Suppose Conditions (3.P1–3.P3) hold. Then we have that (as* $N \to \infty$*),*

(a) The output MAI of the MMSE receiver, $\mathcal{I}_1^{(N)}$*, has a limiting proper complex Gaussian distribution.*

(b) The conditional distribution of the output MAI of the MMSE receiver, given the signatures and received powers, converges in probability to the same proper complex Gaussian distribution as in part (a).

In Theorem 4.1, part (a) shows that the unconditional distribution of $\mathcal{I}_1^{(N)}$ converges (in the weak sense) to a proper complex Gaussian distribution; part (b) establishes that the conditional distribution of $\mathcal{I}_1^{(N)}$, given $\boldsymbol{X}$, converges (in the weak sense) to the same proper complex Gaussian distribution in probability (see also Definition 3.2). We note that the limiting proper complex Gaussian distribution does not depend on the signatures and depends "weakly" on the received powers in the sense that its variance depends on only the empirical distribution of the mean received powers. This indicates that the MMSE receiver is robust to the randomness of the signatures and received powers.

Since the background noise V is proper complex Gaussian, it is clear that conditioned on $\boldsymbol{X}$, the distribution of $\mathcal{I}_2^{(N)}$ converges in probability to a proper complex Gaussian distribution. Also note that conditioned on $\boldsymbol{X}$, $\mathcal{I}_1^{(N)}$ and $\mathcal{I}_2^{(N)}$ are independent. Therefore, based on Thereom 4.1, we conclude that the conditional distribution of the overall output interference, $\mathcal{I}^{(N)}$, given $\boldsymbol{X}$, converges in probability to a proper complex Gaussian distribution. Heuristically, given the signatures and received powers, the output interference in a large system is approximately Gaussian with high probability. This result is particularly useful for systems in which the powers vary at a slower rate than information symbols and the signatures are short (i.e., the period of the signature equals the information symbol period and hence repetition of the same random signatures is adopted). The reasoning is as follows: Assuming the information symbols are independent (which is valid when interleaving and de-interleaving are employed), the output interference, given the signatures and received powers, is independent across symbol intervals, and the Gaussianity of the conditional distribution greatly simplifies the performance analysis and characterization of channel capacity.

We now take a network perspective and proceed to characterize the network capacity of single-class systems.

Building on Theorem 4.1, we take the QoS requirement as meeting the SIR constraints. This is sensible because in view of the Gaussianity of the output interference, the SIR is of fundamental interest for detection and characterization of channel capacity, that is, the SIR is the key parameter that governs the system performance. Also because any scaled version of the MMSE receiver results in the same SIR, it suffices to use any scaled version of the MMSE receiver. Thus, there is really no need for knowledge of the desired user's instantaneous received power for constructing the receiver. This implies that the MMSE receiver is robust to channel uncertainty, which suggests that the MMSE receiver is particularly useful in a multiple-antenna wireless system, where channel estimation becomes more difficult.

In a single-class system, the received powers of all the users are identically distributed. Let F denote the received power distribution and μ its mean. Assume that F is continuous. In what follows, we first characterize the asymptotic SIR, which serves as the basis for identifying the network capacity. Recall that

$$\mathrm{SIR}_1^{(N)} = \frac{P_1(s_1^H M_I^{-1} s_1)^2}{s_1^H M_I^{-1} M_I' M_I^{-1} s_1}.$$

In the special case where the received powers are identically distributed, we have that $M_I = \mu S_1 S_1^H + \eta I$. Note that

$$s_1^H M_I^{-1} M_I' M_I^{-1} s_1 = \sum_{i=2}^{K} |t_i^{(N)}|^2 + \eta s_1^H M_I^{-2} s_1.$$

Let G^* denote the limiting distribution of the eigenvalues of the random matrix $S_1 E_1 S_1^H$ (it has been shown in [10] that the empirical distribution of the eigenvalues of $S_1 E_1 S_1^H$ converges weakly to G^* with probability one; see Appendix B for more details). Using the same techniques as in the proof of [16, Theorem 2.1], it can be shown that

$$\mathrm{SIR}_1^{(N)} \xrightarrow{P} P_1 \int_0^\infty \frac{1}{\lambda + \eta}\, dG^*(\lambda). \tag{6}$$

It then follows that $\mathrm{SIR}_1^{(N)}$ (given P_1) converges in probability to $P_1 \beta'$, where

$$\beta' = \frac{1}{2\mu\eta}\left[\sqrt{(\mu(\alpha - 1) + \eta)^2 + 4\mu\eta} - \mu(\alpha - 1) - \eta\right]. \tag{7}$$

We are now ready to present our main result on the network capacity of a single-class system with MMSE receivers.

Theorem 4.2 *The asymptotic network capacity $\alpha_\infty(a)$ of a system with the MMSE receiver is*

$$\alpha_\infty(a) = \frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu} - \frac{\eta\gamma}{F^{-1}(1-a)}.$$

Roughly speaking, in a large network, if the users choose their signatures randomly and independently, then up to $N \cdot \left(\frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu} - \frac{\eta\gamma}{F^{-1}(1-a)} \right)$ users are admissible in the system.

It is of interest to compare the above result with the corresponding characterization of network capacity of a system with optimal deterministic signatures carried out in [15]. It is shown in [15] that the network capacity of a system with optimal signature allocation is $\frac{F^{-1}(1-a)}{\gamma\mu} + 1 - \frac{\eta}{\mu}$. Compared to Theorem 4.2, we conclude that by allocating signatures "intelligently", the network capacity is increased by $\frac{\eta\gamma}{F^{-1}(1-a)}$. However, it is also clear that as the signal-to-noise ratio (SNR) increases, the gap between the network capacity corresponding to optimal signatures and that corresponding to random signatures vanishes. That is, at the cost of transmission power, we can drive the gap between the network capacity corresponding to optimal signatures and that corresponding to random signatures arbitrarily small. This observation leads to the conclusion that from the viewpoint of network capacity, systems with MMSE receivers are *robust* to the choice of signature sets.

Based on Theorem 4.2, it is clear that there are cases where the network capacity of a system with random spreading is great than one. It is somewhat surprising that "random signatures" can sometimes beat "orthogonal signatures." Mathematically, this is a high-dimensional stochastic geometry problem, and it may be of interest for future investigation.

5 Conclusions

We consider a canonical symbol-synchronous discrete-time model for wireless networks with MMSE receivers. We focus on the cases where the signatures are modeled as random and take values in complex space. First we use the conditional distribution of the output MAI of the MMSE receiver to characterize the system performance. We have found that under the assumptions (3.P1–3.P3), the conditional distribution of the output MAI, given the signatures and received powers, converges in probability to a proper complex Gaussian distribution that does not depend on the signatures and depends weakly on the mean received powers. This result indicates that in a large system, the overall output interference of the MMSE receiver is approximately Gaussian with high probability, and that systems with MMSE receivers are robust to the randomness of the signatures and received powers. Furthermore, because of the Gaussian nature of the output interference, the SIR is of fundamental interest.

Building on the Gaussianity of the output interference, we then take a networking perspective and identify the network capacity of single-class systems with random spreading. We have found that the network capacity can be expressed uniquely in terms of the SIR requirements and received power distributions. Compared to the network capacity corresponding to the optimal signature allocation characterized in [15], we conclude that at the cost of transmission power, the gap between the network capacity corresponding to optimal signatures and that corresponding to random signatures can be made arbitrarily small. This observation leads to the conclusion that from the viewpoint of network capacity, systems with MMSE receivers are robust to the randomness of signatures.

Our results are useful for performance analysis and characterization of system limits. In particular, the bit error probability (corresponding to a specific modulation scheme) can easily be obtained in terms of the SIR. Based on the calculation of the bit error probability, we can calculate packet error probability, which in turn impacts performance measures at the network layer, such as throughput and packet delay. Our result on the network capacity is also potentially useful for network-level resource allocation problems such as admission control and power control in a large system.

References

[1] P. Billingsley, *Probability and Measure.* John Wiley & Sons, Inc., third ed., 1995.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* John Wiley & Sons, Inc., 1991.

[3] E. A. Lee and D. G. Messerschmitt, *Digital Communication.* Kluwer Academic Publishers, second ed., 1994.

[4] U. Madhow and M. L. Honig, "MMSE interference suppression for directed-sequence spread-spectrum CDMA," *IEEE Transactions on Communications*, vol. 42, pp. 3178–3188, Dec. 1994.

[5] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayleigh flat fading," *IEEE Transactions on Information Theory*, vol. 45, pp. 139–157, Jan. 1999.

[6] K. S. Miller, *Complex Stochastic Processes: An Introduction to Theory and Application.* Addison–Wesley, first ed., 1974.

[7] D. Mitra and J. A. Morrison, "A distributed power control algorithm for bursty transmissions on cellular, spread spectrum wireless networks," in *Proc. 5th WINLAB Workshop on Third Generation Wireless Information Networks* (J. M. Holtzman, ed.), pp. 201–212, Kluwer Academic Publishers, 1996.

[8] F. D. Nesser and J. L. Massey, "Proper complex random processes with applications to information theory," *IEEE Transactions on Information Theory*, vol. 39, pp. 1293–1302, July 1993.

[9] H. V. Poor and S. Verdú, "Probability of error in MMSE multiuser detection," *IEEE Transactions on Information Theory*, vol. 43, no. 3, pp. 858–871, May 1997.

[10] J. W. Silverstein and Z. D. Bai, "On the empirical distribution of eigenvalues of a class of large dimensional random matrices," *Journal of Multivariate Analysis*, vol. 54, no. 2, pp. 175–192, 1995.

[11] T. J. Sweeting, "On conditional weak convergence," *Journal of Theoretical Probability*, vol. 2, no. 4, pp. 461–474, 1989.

[12] D. Tse and S. V. Hanly, "Linear multiuser receivers: effective interference, effective bandwidth and user capacity," *IEEE Transactions on Information Theory*, vol. 45, pp. 641–657, Mar. 1999.

[13] S. Verdú, *Multiuser Detection.* Cambridge University Press, 1998.

[14] S. Verdú and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Transactions on Information Theory*, vol. 45, pp. 622–640, Mar. 1999.

[15] J. Zhang and E. K. P. Chong, "CDMA systems in fading channels: Admissibility, network capacity, and power control," *IEEE Transactions on Information Theory*, vol. 46, pp. 962–981, May 2000.

[16] J. Zhang, E. K. P. Chong, and D. N. C. Tse, "Distributions of the output MAI of linear MMSE multiuser receivers in synchronous DS-CDMA systems," in *Proceedings of International Symposium on Information Theory*, (Sorrento, Italy), p. 382, June 2000.

[17] J. Zhang, E. K. P. Chong, and D. N. C. Tse, "Output MAI distributions of linear MMSE multiuser receivers in DS-CDMA systems," *IEEE Transactions on Information Theory*, to appear.

Adaptive Modulation Techniques in Coded CDMA *

Francesco Alesiani   Ezio Biglieri   Giorgio Taricco   Emanuele Viterbo

Dipartimento di Elettronica Politecnico di Torino
Corso Duca degli Abruzzi 24 I-10129 Torino (Italy)
E-mail: `<lastname>@polito.it`

Abstract

In this work we consider the performance of a UMTS CDMA downlink achieved by a single-user (RAKE) and a multiuser (linear MMSE) receiver with adaptive modulation and multicode in the presence of multipath fading. We propose an adaptive transmission scheme that varies the number of virtual users and the modulation spectral efficiency in order to optimize the system throughput. Two multipath channel models are considered with 3 and 9 fading paths. The multiuser MMSE receiver provides a significant performance enhancement with respect to the RAKE receiver with the more favorable multipath channel.

Keywords: CDMA, Spread Spectrum, Adaptive Modulation.

1 Introduction

Several European second-generation mobile communication systems are based on a binary modulation scheme: GMSK. However, to enable future mobile systems to deliver more flexible services, higher spectral efficiencies are called for. To approach the capacity of the transmission channel, its peculiar features have to be taken into account. Thus, since wireless channels exhibit time-variant power, adaptive transmission strategies look attractive. Among these, adaptive modulation (see [5, 10] and references therein) plays a central role, because it increases the data transmission efficiency without increasing the multi-access interference power (other strategies are also possible: for example, in [7, 8] an adaptive-spreading-factor, adaptive-power system is advocated). In their essence, adaptive modulation techniques (AMT) consist of transmitting at the highest possible rate compatible with an assigned quality of service (QoS), as specified by higher-layer requirements such as packet error rate, packet delay, etc. This is obtained by using a hierarchy of different constellations of increasing size. AMT can be combined with a transmit-power control scheme for increased efficiency [10]; however, we do not consider this added feature here.

AMT are based on two steps:

(a) Measurement of the parameters describing the quality of the transmission channel.

---

*This research has been supported by Omnitel Pronto Italia.

(b) Selection of the modulation format based on the optimization of a preassigned cost function.

It should be observed that a basic assumption here is that the channel does not vary too rapidly, otherwise the selection of the modulation format might occur in a situation in which it is badly matched to the channel conditions. Thus, adaptive modulation can only be beneficial in a situation where the Doppler spread is not too wide. This makes AMT especially attractive in an indoor environment, where propagation delays are small and the relative speed between transmitter and receiver is typically low. In these conditions, AMT can work on a frame-by-frame basis.

The simplest propagation parameter that can be used to assess the quality of the transmission channel is the received signal level. In the absence of interferences, this is proportional to the signal-to-noise ratio, and consequently a set of thresholds can be determined such that, when a threshold is crossed upwards, the size of the constellation is increased. Measurement of the signal level is especially easy in a Time Division Duplex (TDD) system: here the signal power in a TDD frame, received at one end of the TDD link, is used to estimate the channel gain in the reverse-link time slot. Otherwise, channel-measurement should be relayed from the receiver to the transmitter.

In the presence of interferences, things become more complex: first, in TDD systems the interferences in uplink and downlink are independent; moreover, a high signal level may correspond to high interference, and consequently it may not be appropriate to increase the constellation size. In these conditions, measurement of the signal-to-noise plus interference ratio (SINR) is more appropriate. Another proposition consists of cancelling the interference prior to signal-level measurement. Here we assume that the transmission powers of all users are the maximum permitted, while the modulation size is selected according to a suitable measure, derived from the SINR, of the QoS.

In this paper, these concepts are applied to UMTS, the 3rd-generation mobile system, where AMT is advocated with the goal of achieving the bit rate of at least 2 Mbit/s over its downlink. Specifically, we consider AMT in conjunction with multicode transmission, where the chip rate and the processing gain are fixed, while the user data bits are split among a number N_{vu} of virtual users, each employing a different signature. Moreover, we examine a situation in which the transmitted data are error-control encoded, so that a QoS measure compatible with coding must be derived. We obtain performance results for the single-user RAKE receiver and the linear multiuser minimum mean-square error (MMSE) receiver.

2 Adaptive modulation

In order to match the transmission rate to the channel conditions, different transmission modes are considered. For a fixed spreading factor L, the transmission modes are defined by the number of virtual users ($N_{\text{vu}} = 1, \dots, L-1$) and the modulation spectral efficiency ($\eta = 2, 4, 6$ bit/s/Hz). The number of virtual users corresponds to the number of spreading codes used in the multicode transmission while the different spectral efficiencies are obtained by using QPSK, 16-QAM and 64-QAM modulation schemes. Note that one spreading code is always reserved for the common pilot channel and its overhead may be reduced by increasing L. Since UMTS is based on orthogonal Hadamard spreading codes, the maximum number of virtual users is $L-1$.

As an example we list in Table 1 the achievable data rates using a rate-$1/3$ turbo code when $L = 4$ in the different modes. The number of data bits per spreading code with $L = 4$ is

1248 when QPSK is used. Furthermore, 16 bits are used to generate pilot symbols, 8 bits for the Transport Format Combination Indicator (TFCI) and 8 bits for the Transmit Power Control (TPC).

	η		
N_{vu}	2 bit/s/Hz	4 bit/s/Hz	6 bit/s/Hz
1	624 kbps	1248 kbps	2496 kbps
2	1248 kbps	2496 kbps	4992 kbps
3	1872 kbps	3744 kbps	7488 kbps

Table 1: *Information bit rates for the different transmission modes.*

Switching between transmission modes represented by the pair (N_{vu}, η) is dictated by the value of the SINR estimated from the pilot channel in the previous slot.

We hasten to observe that additional redundant information must be carried by each frame in order to identify which modulation scheme was used. Hereafter we do not take this into account, and assume that the scheme used is accurately identified by the receiver.

Turbo encoding and decoding. Our simulation results use the rate-1/3 Turbo code specificied in document [13]. For all transmission modes we assumed that the codewords contain the data transmitted by all the virtual users during one slot (0.667 ms). Thus, their length changes according to the transmission mode. This choice has the advantage of producing a fixed delay of one slot for all bit rates and an additional interleaving gain (due to its increased length) when the higher bit rates are selected. The internal interleaver of the turbo code needs to be changed for each trasmission mode. For the previous example we have the interleaver lengths given in Table 2.

	η		
N_{vu}	2 bit/s/Hz	4 bit/s/Hz	6 bit/s/Hz
1	416 bit	832 bit	1664 bit
2	832 bit	1664 bit	3744 bit
3	1248 bit	2496 bit	4992 bit

Table 2: *Interleaver sizes for the different transmission modes.*

In order to combine turbo decoding with higher-order modulations, an additional block called bit-metric computer evaluates the log-likelihood ratios of each bit forming the label of the modulated signal. The decoder uses iterative decoding up to 8 iterations. In order to evaluate and compare the error rates for the different transmission modes in an adaptive environment, we assume that each codeword (slot) carries $N_{\text{vu}} \times \eta/2$ data packets of fixed length. For example with $L = 4$ the transmission mode ($N_{\text{vu}} = 3$, 16-QAM) allows transmission of six data packets. Note that data packets are only *virtual units* of information bits. We thus consider as performance parameters the packet error rate (PER) and the system throughput defined as $R_b(1 - \text{PER})$. Here R_b is the gross information bit rate (which includes the packet errors), given for example in Table 1.

The modulator is implemented according to document [14, sec. 5] with the additional feature of producing multilevel modulated symbols from 16- and 64-QAM constellations. We note that pilot symbols are always trasmitted with QPSK for all transmission modes. The Common

Pilot Channel transmits the complex symbol $(1+j)$ with the same power of the average transmitted power used for the data symbols. For all modulation schemes Gray bit labeling was selected [2].

Spreading is implemented, according to [14, sec. 5] using Orthogonal Variable Spreading Factor (OVSF) codes.

Fast power control. The transmitted signal is power-controlled with 1-dB steps at the slot frequency. The channel level is measured in each slot and a transmit-power UP or DOWN is performed accordingly for the following slot (see [15]). The global effect of the power control is an average increase in the transmitted power which enables to drastically reduce the channel fluctuations due to fading. We recall that in FDD systems the power control command need be transmitted on a feedback channel from mobile to base station, whereas on TDD systems this is not necessary since the transmitter (base station) has a good estimate of the channel level from the reverse link.

Fading channel. The multipath fading channel model was selected among those given in Annex C of [16]. All the results given in this paper were based on the following two power-delay profiles denoted as "Pedestrian A" (good) and "Pedestrian B" (bad). Fig. 1 shows these two profiles.

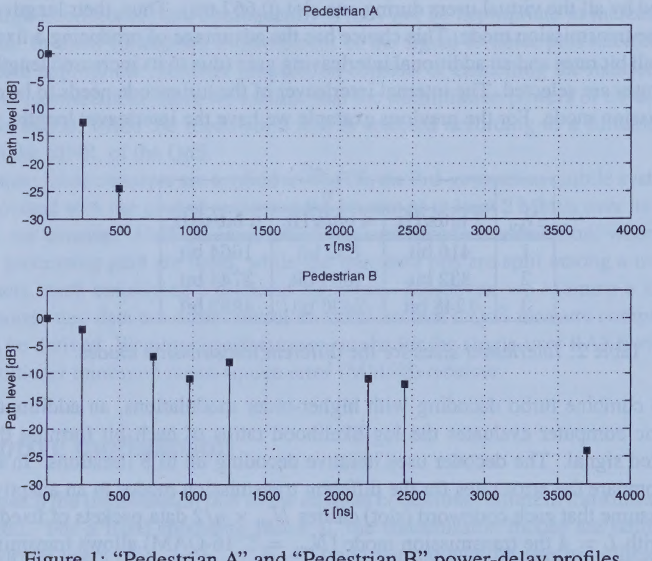


Figure 1: "Pedestrian A" and "Pedestrian B" power-delay profiles.

The time-varying nature of the channel, caused by the relative motion of the receiver with respect to the base station, is modeled according to Jakes' model. In all the following it is assumed that the mobile speed is 3 km/h, corresponding to a Doppler band of 5 Hz. In our simulations, the Jakes Doppler spectrum was approximated by using a three-pole discrete filter.

RAKE receiver. The RAKE receiver we use has as many fingers as the number of paths in the multipath channel and is assumed to have perfect channel state information. It performs maximum-ratio combining of all paths. The output of the RAKE receiver is passed to the despreading filters, which perform as single-user matched filter detectors for all virtual users. Due to the presence of multipath fading, the received signature waveforms are non-orthogonal. Hence, this solution is sub-optimal and its performance depends on N_{vu}.

Linear multiuser MMSE receiver. The discrete-time baseband model of the linear multiuser MMSE receiver for the downlink is described following [6]. We assume that K users (virtual users with multi-code) access a CDMA system with spreading factor L, and that each user transmits a data symbol sequence spanning N symbol intervals represented by the vector

$$\mathbf{d}^{(k)} = (d_1^{(k)}, d_2^{(k)}, \ldots, d_N^{(k)})^T \qquad k = 1, \ldots, K$$

where $(\cdot)^T$ denotes transpose. Symbols are taken from some finite constellation. Let the user signature sequences be defined by the vectors

$$\mathbf{c}^{(k)} = (c_1^{(k)}, c_2^{(k)}, \ldots, c_L^{(k)})^T \qquad k = 1, \ldots, K$$

and the common multipath (P paths) channel impulse response be defined by

$$\mathbf{h} = (h_1, h_2, \ldots, h_P)^T$$

The combined channel-impulse response

$$\mathbf{b}^{(k)} = (b_1^{(k)}, b_2^{(k)}, \ldots, b_{L+P-1}^{(k)})^T \qquad k = 1, \ldots, K$$

is given by the discrete convolution $\mathbf{b}^{(k)} = \mathbf{c}^{(k)} * \mathbf{h}$. We assume that the channel is constant during the transmission of the symbol sequences of length N. Assuming a chip- and symbol-synchronous receiver, the (chip-rate) received sequence is then given by

$$\mathbf{e} = (e_1^{(k)}, e_2^{(k)}, \ldots, e_{NL+P-1}^{(k)})^T = \mathbf{A}\mathbf{d} + \mathbf{n}$$

where $\mathbf{d} = ((\mathbf{d}^{(1)})^T, (\mathbf{d}^{(2)})^T, \ldots, (\mathbf{d}^{(K)})^T)^T$; $\mathbf{n} = (n_1^{(k)}, n_2^{(k)}, \ldots, n_{NL+P-1}^{(k)})^T$ represents the additive noise vector at the receiver which is assumed to be zero-mean complex circularly Gaussian distributed with diagonal covariance matrix $\sigma^2\mathbf{I}$; $\mathbf{A}$ is the $(NL + P - 1) \times NK$ equivalent channel matrix defined by

$$\mathbf{A}((n-1)L + 1 : nL + P, (k-1)N + n) = \mathbf{b}^{(k)} \qquad k = 1, \ldots, K, \ \ n = 1, \ldots, N$$

The matrix colon notation is taken from [4]: $\mathbf{A}(a : b, c)$ denotes the sub-matrix obtained by extracting the elements of $\mathbf{A}$ from row a to b and column c. The MMSE receiver obtains an estimate $\widehat{\mathbf{d}} = \mathbf{H}\mathbf{e}$ of the transmitted data symbols by minimizing the expectation $\mathbb{E}[|\widehat{\mathbf{d}} - \mathbf{d}|^2]$. It can be shown that [6, MMSE-BLE receiver]

$$\widehat{\mathbf{d}} = (\mathbf{A}^\dagger\mathbf{A} + \sigma^2\mathbf{I})^{-1}\mathbf{A}^\dagger\mathbf{e} = \mathbf{A}^\dagger(\mathbf{A}\mathbf{A}^\dagger + \sigma^2\mathbf{I})^{-1}\mathbf{e} \tag{1}$$

The vector $\widehat{\mathbf{d}}$ is then passed to the bit-metric computer for the turbo decoder.

Complexity issues. The complexity of the MMSE receiver depends on NK, the size of the matrix $\mathbf{A}^\dagger\mathbf{A}$. Actually, equation (1) represents the solution of a linear equation system, whose complexity is known to be $O((NK)^3)$ [4]. Since the UMTS system assumes 2560 chips/slot, with spreading factor $L = 4$ and $K = 3$ virtual users, we need to invert a matrix of size $NK = 1920$. However, the received sequence of length $NL + P - 1$ can be partitioned into (N/N_s) shorter subsequences of length $N_sL + P - 1$ defined as $\mathbf{e}_i = \mathbf{e}((i-1)N_sL + 1 : iN_sL + P)$ for $i = 1, \ldots, N/N_s$ (it is assumed that N_s divides N). The resulting reduction in complexity is given by the factor $(N/N_s)^3$. In our simulation results we assumed $N_s = 4$, which was found to give a good trade-off between complexity and performance results.

Signal-to-noise ratios definitions. The system performance is obtained for different values of E_b/N_0. E_b is the received energy per information bit. N_0 is the sum of the one-sided power spectral density of the interference from other cells, the interchannel interference and the thermal noise. Note that N_0 does not include the multiple-access interference (MAI) which depends on the value of N_{vu} selected for the transmission mode. The overall SINR is measured by E_s/I_0 where E_s is the received energy per modulation symbol and I_0 is given by the sum of N_0 and the MAI. We can also write $E_b = E_s/(R_c R_d)$ where R_c is the code rate defined as

$$R_c = \frac{\text{N. of information bits}}{\text{N. of coded bits and tail bits}}$$

and R_d accounts for the rate loss due to the control channel bits overhead in the slots.

3 Adaptive selection of transmission modes

In order to select automatically and adaptively the transmission modes, the receiver estimates the SINR from the common pilot channel and compares it against a set of thresholds. In this section we discuss how to determine this set.

In [11] channel coding is not considered and thresholds are chosen by optimizing a cost function which combines QoS, expressed by an upper bound to bit error probability, with transmission efficiency, expressed by the number of transmitted bits per symbol. In this study we use a different approach based on the knowledge of the occurrence of decoding errors which can be reliably detected by checking the data cyclic redundancy check (CRC) symbols.

Let E denote the random variable defining the occurrence of a decoding error taking values 0 when the received codeword is correct and 1 otherwise. Let γ denote the received SINR and $p(\gamma, E = e)$ be the joint probability density function (pdf) of the SINR and the decoding outcome $e = 0, 1$. These pdf's depend on the transmission mode and on the value of E_b/N_0. They can be estimated for the given channel during a training period. We select a number M of transmission modes and denote each of them by an index $\mu = 1, \ldots, M$. Transmission mode μ corresponds to the pdf $p_\mu(\gamma, E = e)$ and a the bit rate $R_{b,\mu}$. Since our goal is to achieve reliable transmission at the maximum bit rate, a maximum a-posteriori (MAP) adaptive receiver can be operated by selecting the mode

$$\widehat{\mu}(\gamma) = \arg \max_{\mu : p_\mu(\gamma, E=0) > p_\mu(\gamma, E=1)} R_{b,\mu}$$

This rule translates into a set of thresholds delimiting the intervals where $\widehat{\mu}(\gamma)$ takes on a fixed value. As an example, we consider the UMTS RAKE receiver with spreading factor $L = 4$ and the following transmission modes:

1. $N_{\text{vu}} = 2$, QPSK
2. $N_{\text{vu}} = 3$, QPSK
3. $N_{\text{vu}} = 3$, 16-QAM
4. $N_{\text{vu}} = 3$, 64-QAM

In this case, the thresholds obtained are approximately -5, 1, and 4 dB for values of E_b/N_0 ranging from 6 to 10 dB. For different values of E_b/N_0, a slight change is expected but, in our simulations, we kept these values fixed.

4 Results

In this section we report simulation results relevant to the downlink of a UMTS CDMA system with a rate-1/3 turbo-code defined according to [13], spreading factor $L = 4$, chip rate of 3.84 Mchip/s, slot duration of 10 ms (each slot carrying one codeword), fast power control in 1-dB steps, and Walsh-Hadamard spreading codes. We consider a multipath frequency-selective Rayleigh fading channel described by the Pedestrian A/B models whose power-delay profiles are given in Figure 1.

Figures 2 and 3 report the results obtained by using a single-user RAKE receiver and a multiuser MMSE receiver, respectively. Each figure contains four plots: those at left refer to the Pedestrian A channel model; those at righ refer to the Pedestrian B model; the upper ones show the throughput and the lower ones report the PER. Simulation runs are 10 seconds long (i.e., 1000 slots). Power control increases the average value of E_b/N_0 by 4.5 dB.

Each diagram contains results referring to nine fixed transmission modes (N_{vu}, η) for $N_{\text{vu}} = 1, 2, 3$ and $\eta = 1, 2, 3$ and the adaptive transmission scheme using the transmission modes $(N_{\text{vu}}, \eta) = (2, 2)$, (3,2), (3,4), and (3,6).

The results show an overall performance enhancement of the multiuser MMSE receiver that is particularly noticeable with the Pedestrian A channel model. With the Pedestrian B model the improvement is limited since the most demanding modes are not achievable. However, we note a considerable reduction of the PER with the MMSE receiver.

It can be seen that the performance of the adaptive receiver is better than the performance of any fixed mode receiver over the whole range of E_b/N_0. Additionally, the adaptive receiver performance degrades smoothly while the E_b/N_0 is reduced in contrast with the abrupt degradation characteristic of fixed-mode receivers.

5 Conclusions

In this work we investigated the performance of downlink transmission at 2 Mbit/s over a UMTS CDMA link with single-user (RAKE) and multiuser (linear MMSE) reception and adaptive modulation and multicode. The multiuser MMSE receiver significantly reduces the PER and marginally increases the throughput with respect to the RAKE receiver. The MMSE receiver is among the simplest multiuser detectors, and hence further improvements may be expected with other multiuser detectors closer to optimum performance, albeit at the price of an increased complexity.

In summary, we showed that implementing an adaptive transmission policy in the UMTS downlink provides both performance enhancement and smoother response to variations of the interference level.

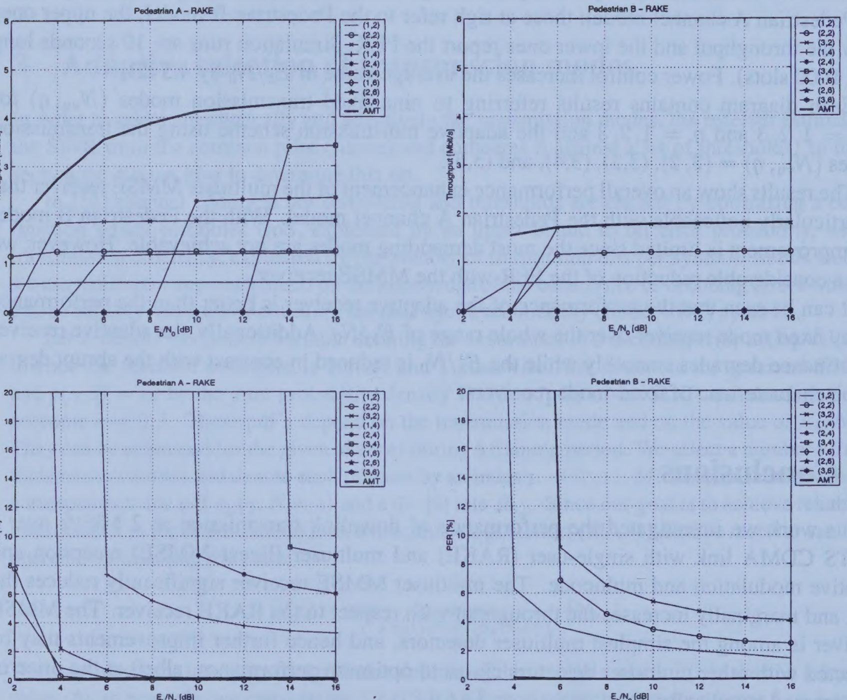


Figure 2: Simulation results (throughput and PER) for the single-user RAKE receiver.

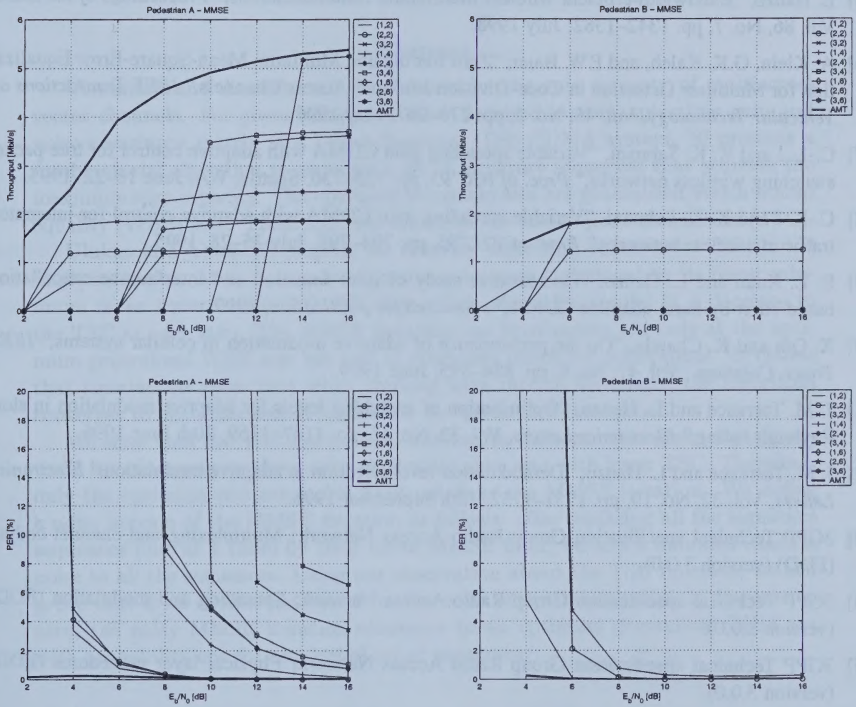


Figure 3: Simulation results (throughput and PER) for the multiuser MMSE receiver.

References

[1] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*. New York: Plenum/Kluwer, 1999.

[2] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," em IEEE Trans. Inform. Theory, Vol. 44, No. 3, pp. 927–946, May 1998.

[3] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, Vol. 45, No. 10, pp. 1218–1230, October 1997.

[4] G.H. Golub and C.F. Van Loan, *Matrix Computations 3rd edition*. London: Johns Hopkins, 1996.

[5] L. Hanzo, "Bandwidth-efficient wireless multimedia communications," *Proceedings of the IEEE*, Vol. 86, No. 7, pp. 1342–1382, July 1998.

[6] A. Klein, G.K. Kaleh, and P.W. Baier, "Zero forcing and Minimum Mean-Square-Error Equalization for Multiuser Detection in Code-Division Multiple-Access Channels," *IEEE Transactions on Vehicular Technology*, Vol. 45, No. 2, pp. 276–287, May 1996.

[7] C.-L. I and K. K. Sabnani, "Variable spreading gain CDMA with adaptive control for true packet switching wireless networks," *Proc. of ICC'95*, pp. 725–730, Seattle, WA, June 18–22, 1995.

[8] C.-L. I and K. K. Sabnani, "Variable spreading gain CDMA with adaptive control for integrated traffic in wireless networks," *Proc. of VTC'95*, pp. 794–798, July 25–28, 1995.

[9] E. L. Kuan and L. Hanzo, "Comparative study of joint-detection and interference cancellation based burst-by-burst adaptive CDMA," *submitted for publication*, 1999.

[10] X. Qiu and K. Chawla, "On the performance of adaptive modulation in cellular systems," *IEEE Trans. Commun.*, Vol. 47, No. 6, pp. 884–895, June 1999.

[11] J. M. Torrance and L. Hanzo, "Optimisation of switching levels for adaptive modulation in slow Rayleigh fading," *Electronics Letters*, Vol. 32, No. 13, pp. 1167–1169, 20th June 1996.

[12] J. M. Torrance and L. Hanzo, "Demodulation level selection in adaptive modulation," *Electronics Letters*, Vol. 32, No. 19, pp. 1751–1752, 12th September 1996.

[13] 3GPP Technical specification Group Radio Access Network; Multiplexing and channel coding (FDD) (version 3.0.0).

[14] 3GPP Technical specification Group Radio Access Network; Spreading and modulation (FDD) (version 3.0.0).

[15] 3GPP Technical specification Group Radio Access Network; Physical layer procedures (FDD) (version 3.0.0).

[16] Tdoc ETSI SMG2 351/98, RTT Revision - Performance Results (Annex C – Capacity and coverage results).

Ensuring convergence of the MMSE iteration for interference avoidance to the global optimum *

Pablo Anigstein and Venkat Anantharam
Dept. of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, CA 94720
{pabloa,ananth}@eecs.berkeley.edu

Abstract

Viswanath and Anantharam [8] characterize the sum capacity of multiaccess vector channels. For given number of users, received powers, spreading gain and noise covariance matrix in a Direct-Sequence (DS) CDMA system, [8] presents a combinatorial algorithm to generate a set of signature sequences that achieves the maximum sum capacity. The optimum sequences sets are generalized Welch bound equality (WBE) sets. These sets also minimize the total square correlation (TSC).

Ulukus and Yates [5] propose an iterative algorithm suitable for distributed implementation: at each step one signature sequence is replaced by its linear minimum mean square error (MMSE) filter. This algorithm results in a decrease in the TSC at each step. The MMSE iteration has fixed points not only at the optimum generalized WBE sets but also at other sets which are suboptimal. [5] claims that simulations show that when starting with randon sequences, the algorithm converges to optimum sets of sequences, but gives no formal proof.

We show that the TSC has no local minima, in the sense that given any suboptimal set of sequences, there exist arbitrarily close sets with lower TSC. Therefore, only the optimum sets are stable fixed points of the MMSE iteration. We define a noisy version of the MMSE iteration as follows: after replacing all the signature sequences (one at a time) by their linear MMSE filter, we add a bounded random noise to all the sequences. Using our observation about the TSC function, we can prove that if we choose the bound on the noise adequately, making it decrease to zero, the noisy MMSE iteration converges to an optimum generalized WBE set with probability one for any initial set of sequences.

1 Introduction

We consider a synchronous code-division multiple-access (CDMA) system with K users and processing gain N. The received signal at the base station over one symbol interval will be represented by an N-dimensional column vector y:

$$y = \sum_{k=1}^{K} \sqrt{p_k} x_k s_k + n \tag{1}$$

---

*This research is supported by EPRI/DOD Complex Interactive Networks under contract number EPRI-W08333-04, NSF ANI 9872764, ECS 9873086 and IIS 9941569

Here s_k is an N-dimensional column vector corresponding to the signature sequence of user k, assumed unit-energy (i.e. $s_k^T s_k = 1$). The power received from user k is p_k. The information transmitted by user k is modeled by the random variable x_k having zero mean and unit variance. The noise is assumed Gaussian zero-mean with covariance $E[nn^T] = W$, a $K \times K$ symmetric positive definite matrix.

If we write $S = \begin{bmatrix} s_1 & \dots & s_K \end{bmatrix}$, $D = \text{diag}(p_1, \dots, p_K)$ and $x = \begin{bmatrix} x_1 & \dots & x_K \end{bmatrix}^T$ equation (1) can be rewritten as

$$y = SD^{1/2}x + n \tag{2}$$

We assume N, K, p_k ($k \in \{1, \dots, K\}$) and W are given and fixed. Thus a configuration is determined by the signatures $S \in \mathcal{S}$ where

$$\mathcal{S} \triangleq \{\begin{bmatrix} a_1 & \dots & a_K \end{bmatrix} : a_k \in \mathbb{S}^{N-1} \; \forall k \in \{1, \dots, K\}\} \tag{3}$$

with $\mathbb{S}^{N-1} = \{x \in \mathbb{R}^N : \|x\| = 1\}$ the unit-sphere in $\mathbb{R}^N$.

The sum capacity of channel 2 is [8]

$$C_{sum}(S) = \frac{1}{2} \log \det \left(I + W^{-1}SDS^T\right) = \frac{1}{2} \log \det \left(SDS^T + W\right) - \frac{1}{2} \log \det(W) \tag{4}$$

The problem of maximizing $C_{sum}(S)$ over all $S \in \mathcal{S}$ and finding an optimum signature sequence configuration is solved in [8], where majorization theory is used [2]. The sum capacity is a Schur-concave function of the eigenvalues of $SDS^T + W$. The set of vectors of eigenvalues of $SDS^T + W$ as S varies in $\mathcal{S}$ has a Schur-minimal element [9]. Thus this element corresponds to configurations S that achieve the maximum of C_{sum}.

We define a generalized total square correlation (TSC) function as [4]

$$\text{TSC}(S) = \text{Trace}\left[\left(SDS^T + W\right)^2\right] \tag{5}$$

The TSC is a Schur-convex function of the eigenvalues of $SDS^T + W$. Thus the configurations which achieve the maximum of C_{sum} and those which achieve the minimum of TSC are the same.

Iterative algorithms aimed to minimize the TSC are proposed in [5] and [3]. These algorithms update one signature sequence at a time and are amenable to distributed implementation. The TSC at each iteration is non-increasing, but no formal proof of convergence to the minimum of TSC in the general case is given. A modified algorithm is proposed in [4] to prove convergence, but it has increased complexity and is not suitable for distributed implementation.

We consider the MMSE update algorithm [5]. In the next section we define the MMSE update, state some properties and characterize the fixed configurations. In section 3 we observe and sketch a proof that TSC has no minima other than the global minima. Motivated by this result, in section 4 we define a modified version of the MMSE update adding noise. We prove that if the noise bound is chosen adequately, the noisy MMSE update converges to the optimum TSC almost surely regardless of the initial configuration.

2 MMSE update

Given S and $k \in \{1, \dots, K\}$ we will write $D_k = diag(p_1, \dots, p_{k-1}, p_{k+1}, \dots, p_K)$ and $S_k = \begin{bmatrix} s_1 & \dots & s_{k-1} & s_{k+1} & \dots & s_K \end{bmatrix}$.

The MMSE unit-norm linear filter for user k is

$$c_k(S) = \frac{1}{\sqrt{s_k^T \left(S_k D_k S_k^T + W\right)^{-2} s_k}} \left(S_k D_k S_k^T + W\right)^{-1} s_k \tag{6}$$

or equivalently

$$c_k(S) = \frac{1}{\sqrt{s_k^T \left(SDS^T + W\right)^{-2} s_k}} \left(SDS^T + W\right)^{-1} s_k \tag{7}$$

The filter $c_k(S)$ can be shown to minimize the signal-to-interference ratio (SIR) of user k over all linear receivers (for a derivation in the case of white noise $W = wI$ see section 6.2 of [6]).

We define the MMSE user k update function as

$$\Phi_k(S) = \left[\ s_1 \ \ \ldots \ \ s_{k-1} \ \ c_k(S) \ \ s_{k+1} \ \ \ldots \ \ s_K\ \right] \tag{8}$$

which replaces the signature sequence for user k by the corresponding normalized linear MMSE filter. The following lemma [5] states that this update strictly decreases the TSC except when the signature sequence for user k coincides with the MMSE filter.

Lemma 1

$$\forall S \in \mathcal{S}: \ \mathrm{TSC}(\Phi_k(S)) \leq \mathrm{TSC}(S), \quad \textit{with equality iff } s_k = c_k(S) \tag{9}$$

Proof : See [5] ([5] considers the case of white noise ($W = wI$) and equal received powers, but the proof holds for arbitrary W and powers).

□

Consider the MMSE update dynamics in $\mathcal{S}$:

$$S^{(n+1)} = \Phi_{n+1}(S^{(n)}) \tag{10}$$

where we define Φ_n for $n > K$ setting $\Phi_n = \Phi_{n-K}$. This corresponds to replacing each signature sequence using the MMSE update, one at a time. The signature sequences are updated in a deterministic order. This assumption could be relaxed, and with some more work the results could be extended under suitable conditions to partially asynchronous updates [1].

Note that given any initial configuration $S^{(0)} \in \mathcal{S}$, the sequence $\mathrm{TSC}(S^{(n)})$ defined by equation (10) converges because it is non-increasing by lemma 1 and bounded below.

The MMSE update function is defined as

$$\Phi(S) = \Phi_K(\Phi_{K-1}(\ldots \Phi_1(S))) \tag{11}$$

By lemma 1 we obtain that

$$\mathrm{TSC}(\Phi(S)) \leq \mathrm{TSC}(S), \quad \text{with equality iff } \Phi(S) = S \tag{12}$$

Let F_Φ be the set of fixed configurations of Φ:

$$F_\Phi = \{S \in \mathcal{S} : \Phi(S) = S\} \tag{13}$$

We summarize some properties of the fixed configurations of the MMSE update.

Lemma 2 *If $S = \left[\, s_1 \;\; \dots \;\; s_K \,\right] \in F_\Phi$ then*

(a) For all $k \in \{1, \dots, K\}$, s_k is an eigenvector of $SDS^T + W$.

(b) There exists an orthonormal basis of (common) eigenvectors of SDS^T and W. Equivalently, matrices SDS^T and W commute.

(c) Let n be the number of distinct eigenvalues $\mu_1, \dots, \mu_n$ of $SDS^T + W$. There exist a partition $J_1, \dots, J_n$ (with possibly some of the J_i empty) of the set $\{1, \dots, K\}$ and a partition $I_1, \dots, I_n$ of the set $\{1, \dots, N\}$ such that for all $i \in \{1, \dots, n\}$:

$$\begin{aligned} (SDS^T + W)s_k &= \mu_i s_k \quad \forall k \in J_i && (14)\\ s_k^T s_l &= 0 \quad \forall k \in J_i,\ l \in J_j,\ j \neq i && (15)\\ \mu_i &= \frac{\sum_{k \in J_i} p_k + \sum_{m \in I_i} w_m}{|I_i|} && (16)\\ \mathrm{TSC}(S) &= \sum_{i=1}^{n} \frac{\left(\sum_{k \in J_i} p_k + \sum_{m \in I_i} w_m\right)^2}{|I_i|} && (17) \end{aligned}$$

where $|I_i|$ is the number of elements in I_i and $w_1, \dots, w_N$ are the eigenvalues of W. Furthermore, $|I_i|$ is equal to the multiplicity of μ_i.

Proof : Apply lemma 1 to obtain $s_k = c_k(S)$ and for all $k \in \{1, \dots, K\}$ and thus

$$(SDS^T + W)\, s_k = \lambda_k s_k \tag{18}$$

where $\lambda_k = \left[s_k^T \left(SDS^T + W\right)^{-2} s_k \right]^{-\frac{1}{2}}$ and (a) follows.

To prove (b) and (c) let n be the number of distinct eigenvalues $\mu_1, \dots, \mu_n$ of $SDS^T + W$. From (a) all s_k are eigenvectors of $SDS^T + W$, so we can partition the set $\{1, \dots, K\}$ grouping the signatures associated to the same eigenvalues. I.e. if we define

$$J_i \triangleq \{k \in \{1, \dots, K\} : (SDS^T + W)s_k = \mu_i s_k\} \tag{19}$$

the J_i are disjoint and $\bigcup_{i=1}^{n} J_i = \{1, \dots, K\}$. As $SDS^T + W$ is a symmetric matrix, eigenvectors associated with distinct eigenvalues are orthogonal and (15) is proved. If we write $S_{J_i} = [s_k, k \in J_i]$ and $D_{J_i} = \mathrm{diag}\,(p_k, k \in J_i)$ it follows

$$(S_{J_i} D_{J_i} S_{J_i}^T + W)s_k = \mu_i s_k \quad \forall k \in J_i, \quad \forall i \in \{1, \dots, n\} \tag{20}$$

Thus we get

$$(S_{J_i} D_{J_i} S_{J_i}^T + W) S_{J_i} D_{J_i} S_{J_i}^T = \mu_i S_{J_i} D_{J_i} S_{J_i}^T \quad \forall i \in \{1, \dots, n\}$$

which implies that W and $S_{J_i} D_{J_i} S_{J_i}^T$ commute. As $SDS^T = \sum_{i=1}^{n} S_{J_i} D_{J_i} S_{J_i}^T$ we get that W and SDS^T commute. Therefore there exists an orthonormal basis $q_1, \dots, q_N$ of eigenvectors of W and SDS^T. Thus $q_1, \dots, q_N$ are eigenvectors of $SDS^T + W$. W.l.o.g. we assume $Wq_j = w_j q_j$. Then (16) and (17) are obtained by choosing the partition of the set $\{1, \dots, N\}$ as follows:

$$I_i \triangleq \{j \in \{1, \dots, N\} : \ (SDS^T + W)\, q_j = \mu_i q_j\}$$

Note that condition (a) in the lemma above is equivalent to S being a fixed configuration because from (7) $c_k(S) = s_k$ iff s_k is an eigenvector of $SDS^T + W$:

$$S \in F_\Phi \Longleftrightarrow \forall k \in \{1, \ldots, K\} : s_k \text{ is an eigenvector of } SDS^T + W \tag{21}$$

By condition (c), if S is a fixed configuration we can partition the signatures s_k in mutually orthogonal sets each associated with a distinct eigenvalue of $SDS^T + W$.

Given $S^{(0)} \in \mathcal{S}$ we can define the ω-limit set with respect to the dynamics (10) as:

$$\omega_\Phi(S^{(0)}) = \{S \in \mathcal{S} : \exists n_1 < n_2 < \ldots \text{ s.t. } \lim_{m\to\infty} S^{(n_m)} = S\} \tag{22}$$

The following lemma shows that for any initial set of signature sequences, the MMSE update (10) converges to the set of fixed configurations.

Lemma 3 *Given any* $S^{(0)} \in \mathcal{S}$,

$$\omega_\Phi(S^{(0)}) \subset F_\Phi \tag{23}$$

Proof : If $S \in \omega_\Phi(S^{(0)})$ then $\exists m_1 < m_2 < \ldots$ s.t. $\lim_{l\to\infty} S^{(m_l)} = S$. For some $k \in \{1, \ldots, K\}$, m_l is a multiple of k for infinitely many l, let n_l be the corresponding subsequence. Then $S^{(n_l)} \to S$ as $l \to \infty$. By continuity of Φ_{k+1}, $\Phi_{k+1}(S^{(n_l)}) \to \Phi_{k+1}(S)$ as $l \to \infty$. Now assume $\Phi_{k+1}(S) \neq S$. Then by lemma 1, $\mathrm{TSC}(\Phi_{k+1}(S)) < \mathrm{TSC}(S)$. Let $\Delta = \mathrm{TSC}(S) - \mathrm{TSC}(\Phi_{k+1}(S))$. Then, as TSC is continuous, there exists p such that $\forall l > p$ it is $\mathrm{TSC}(S^{(n_l+1)}) < \mathrm{TSC}(S^{(n_l)}) - \frac{\Delta}{2}$. Thus $\mathrm{TSC}(S^{(n_{l+1})}) < \mathrm{TSC}(S^{(n_l)}) - \frac{\Delta}{2}$ for $p > l$ and therefore $\mathrm{TSC}(S^{(n_l)}) \to -\infty$ as $l \to \infty$. This is a contradiction because TSC is positive, and thus $\Phi_{k+1}(S) = S$. But then $\Phi_{k+1}(S^{(n_l)}) = S^{(n_l+1)} \to \Phi_{k+1}(S) = S$ as $l \to \infty$. Recurring to the same argument as before we now get $\Phi_{k+2}(S) = S$. Repeating this argument $(K-2)$ more times we get $\Phi(S) = S$ as we wanted to prove.

□

We conclude that for any initial condition the MMSE update approaches the set of fixed configurations as $n \to \infty$. As TSC is a continuous function, this implies that $\lim_{n\to\infty} \mathrm{TSC}(S^{(n)}) \in T_F$ where

$$T_F = \{\mathrm{TSC}(S) : S \in F_\Phi\} \tag{24}$$

Note that from lemma 2(c), T_F has a finite number of elements because there is a finite number of ways of partitioning the sets $\{1, \ldots, K\}$ and $\{1, \ldots, N\}$. Let τ be the minimum of the TSC:

$$\tau \triangleq \min_{S\in\mathcal{S}} \mathrm{TSC}(S) \tag{25}$$

$\mathcal{S}$ is a compact set, so the minimum is attained and we can define the set of optimal configurations:

$$\Omega \triangleq \{S \in \mathcal{S} : \mathrm{TSC}(S) = \tau\} \tag{26}$$

Clearly we have $\Omega \subset F_\Phi$, because if $S \in \Omega$ then $\tau = \mathrm{TSC}(S) \geq \mathrm{TSC}(\Phi(S)) \geq \mathrm{TSC}(S)$ and therefore $\mathrm{TSC}(\Phi(S)) = \tau$ which implies by (12) that $\Phi(S) = S$. But it is easy to see that F_Φ contains non-optimal configurations, i.e. $F_\Phi \neq \Omega$ except for the trivial case $N = 1$. Therefore T_F has more than one element and we cannot conclude that $\lim_{n\to\infty} \mathrm{TSC}(S^{(n)}) = \tau$ as we would like. Simulations suggest that if the initial condition $S^{(0)}$ is chosen randomly, then $\mathrm{TSC}(S^{(n)})$ converges to τ with probability one [5], but no formal proof has been given.

3 Minima of TSC

An important property of the TSC function is that it has no local minima other than the global minima. To state this formally, let us first define a metric on $\mathcal{S}$. Given $S, S' \in \mathcal{S}$, let

$$d(S, S') = \max_{k=1 \ldots K} \arccos(s_k^T s_k') \tag{27}$$

Note that the triangle inequality holds:

$$\begin{aligned} d(S, S'') = & \max_{k=1 \ldots K} \arccos(s_k^T s_k'') \leq \max_{k=1 \ldots K} \left[\arccos(s_k^T s_k') + \arccos(s'^T_k s''_k)\right] \\ \leq & \max_{k=1 \ldots K} \arccos(s_k^T s_k') + \max_{k=1 \ldots K} \arccos(s'^T_k s''_k) = d(S, S') + d(S', S'') \end{aligned}$$

and d is a metric. Given $S \in \mathcal{S}$ and $\theta \in (0, \pi]$ let $B[S, \theta]$ be the closed ball of radius θ centered at S:

$$B[S, \theta] = \{S' \in \mathcal{S} : d(S', S) \leq \theta\} \tag{28}$$

Lemma 4 *Local minima of TSC are global, i.e. if $S \notin \Omega$, then for all $\epsilon \in (0, \pi]$ there exists $S' \in B[S, \epsilon]$ with* $\mathrm{TSC}(S') < \mathrm{TSC}(S)$.

Proof : Here we sketch a proof for the case of white noise, i.e. $W = wI$. This can be extended for the case of colored noise [1].

Take any $S \in \mathcal{S}$. Consider two cases:

1. Assume $\exists k \in \{1, \ldots, K\}$ such that s_k is not an eigenvector associated with the minimum eigenvalue of $(S_k D_k S_k^T + W)$. Let λ be the minimum eigenvalue of $(S_k D_k S_k^T + W)$ and let v be a unit-norm eigenvector associated with λ. Take S' with $s_j' = s_j$ for $j \neq k$ and $s_k' = \alpha s_k + \beta v$, where $\alpha = \cos\left(\frac{\epsilon}{2}\right)$ and $\beta = -\alpha s_k^T v + \mathrm{sign}(s_k^T v)\sqrt{\alpha^2 \left(s_k^T v\right)^2 + 1 - \alpha^2}$. This is valid because $s'^T_k s_k' = \alpha^2 + \beta^2 + 2\alpha\beta s_k^T v = 1$. We see that $\beta s_k^T v \geq 0$ and thus $s_k^T s_k' = \alpha + \beta s_k^T v \geq \alpha$ and $d(S, S') = \arccos\left(s_k^T s_k'\right) \leq \frac{\epsilon}{2}$. Direct computation shows:

   $$\mathrm{TSC}(S) - \mathrm{TSC}(S') = 2p_k(1 - \alpha^2)\left[s_k^T \left(S_k D_k S_k^T + W\right) s_k - \lambda\right] \tag{29}$$

   As $\left(S_k D_k S_k^T + W\right)$ is a symmetric matrix with minimum eigenvalue λ and s_k is not an eigenvector associated with λ, $s_k^T \left(S_k D_k S_k^T + W\right) s_k > \lambda$. Also $1 - \alpha^2 = \sin^2\left(\frac{\epsilon}{2}\right) > 0$ and therefore $\mathrm{TSC}(S) > \mathrm{TSC}(S')$.

2. Assume $\forall k \in \{1, \ldots, K\}$, s_k is an eigenvector associated with the minimum eigenvalue of $(S_k D_k S_k^T + W)$. Then S is a fixed configuration, and we can partition the K users and the N degrees of freedom as in lemma 2(c). W.l.o.g. assume the eigenvalues of $SDS^T + W$ are ordered $\mu_1 > \mu_2 > \ldots > \mu_n$. Let v be an eigenvenctor of $SDS^T + W$ associated with μ_n.

   - Assume there exists $i \in \{1, \ldots, n-1\}$ with $|J_i| > |I_i|$. Take S' with $s_k' = s_k$ for $k \notin J_i$ and $s_k' = \cos(\alpha_k)s_k + \sin(\alpha_k)v$ for $k \in J_i$, where α_k are chosen so that the column vector $\alpha \in \mathbb{R}^{|J_i|}$ with elements $(\alpha_k,\ k \in J_i)$ satisfies $S_{J_i} D_{J_i} \alpha = 0$ (this can be done because S_{J_i} has rank $|I_i|$). With this choice, after some manipulation we get:

     $$\mathrm{TSC}(S) - \mathrm{TSC}(S') = 2(\mu_i - \mu_n)\|D_{J_i}\alpha\|^2 + o(\|\alpha\|^4) \tag{30}$$

     So it suffices to make α small enough to get $\mathrm{TSC}(S') < \mathrm{TSC}(S)$ and $d(S, S') = \max_{k \in J_i} \alpha_k \leq \epsilon$.

- If $|J_i| \leq |I_i|$ for all $i \in \{1, \dots, n-1\}$ it can be proved that, in the sense defined in [7], all the users in $\bigcup_{i=1}^{n-1} J_i$ are oversized, and in fact S attains the optimum TSC, i.e. $\mathrm{TSC}(S) = \tau$. To see this note that from (16) we have

  $$\mu_i = \frac{\sum_{k \in J_i} p_k}{|I_i|} + w \tag{31}$$

  From now on consider $i \in \{1, \dots, n-1\}$. As $\mu_i > \mu_n$ it can be shown $|J_i| = |I_i|$ (because, when $W = wI$, $|J_i| < |I_i|$ is only possible if $|J_i| = 0$. But then it would be $\mu_i = w \leq \mu_n$). Using (20), for all $k \in J_i$,

  $$s_k^T \left(S_{J_i} D_{J_i} S_{J_i}^T\right) s_k = p_k + \sum_{l \in J_i \setminus \{k\}} p_l \left(s_k^T s_l\right)^2 = \mu_i - w \geq p_k \tag{32}$$

  Then from (31), for all $k \in J_i$ we have $\sum_{l \in J_i} p_l \geq |J_i| p_k$ and therefore $p_k = p_l$ and $s_k^T s_l = 0$ for all $k, l \in J_i$ with $k \neq l$. So we can define p_{J_i} by $p_{J_i} = p_k$ for $k \in J_i$ and using (31) $\mu_i = p_{J_i} + w$. We get $p_{J_1} > p_{J_2} > \dots > p_{J_{n-1}} > \frac{\sum_{k \in J_n} p_k}{|I_n|}$. This implies that the users in $\bigcup_{i=1}^{n-1} J_i$ are oversized. But each of this users has a signature that is orthogonal to all the other signatures, and thus it is easy to see that S is an optimal signature allocation [7], i.e. $\mathrm{TSC}(S) = \tau$.

□

We observe that lemma 4 implies that all the non-optimal fixed configurations are unstable, in the sense that given $S \in F_\Phi \backslash \Omega$, for all $\epsilon > 0$ there exists $S' \in B[S, \epsilon]$ such that for the MMSE update with $S^{(0)} = S'$ we have $\lim_{n \to \infty} \mathrm{TSC}(S^{(n)}) \leq \mathrm{TSC}(S') < \mathrm{TSC}(S)$. On the other hand, for $S \in \Omega$ there exists $\epsilon > 0$ such that for all $S' \in B[S, \epsilon]$ the MMSE update with $S^{(0)} = S'$ satisfies $\lim_{n \to \infty} \mathrm{TSC}(S^{(n)}) = \tau$. This last assertion follows from the fact that T_F is finite and TSC is continuous.

4 Noisy MMSE update

Our last observation on the TSC is key to understand the convergence of the MMSE update. We will next slightly modify the MMSE update algorithm adding noise. To this end we first make some definitions. Given two unit-norm orthogonal vectors v_1, v_2 ($v_1, v_2 \in \mathbb{S}^{N-1}$ with $v_1^T v_2 = 0$) and an angle θ, let $h(v_1, v_2, \theta)$ denote the rotation of v_1 of angle θ towards v_2:

$$h(v_1, v_2, \theta) \triangleq \cos\theta v_1 + \sin\theta v_2 \tag{33}$$

Analogously, given $S, R \in \mathcal{S}$ with $s_k^T r_k = 0$ and $\theta \in \mathbb{R}^K$ let

$$h(S, R, \theta) \triangleq \left[\ h(s_1, r_1, \theta_1) \ \ \dots \ \ h(s_K, r_K, \theta_K)\ \right]$$

Given a sequence of angles $\{\theta_{max}^{(n)}\} \subset (0, 2\pi)$, we define the MMSE noisy iteration as:

$$S^{(n+1)} = h(\Phi(S^{(n)}), R^{(n+1)}, \theta^{(n+1)}) \tag{34}$$

where $r_k^{(n)}, \theta_k^{(n)}$, $k \in \{1, \dots, K\}$, $n \geq 1$ are independent random variables, $\theta_k^{(n)}$ is uniform $(0, \theta_{max}^{(n)})$ and $r_k^{(n)}$ is a random (i.e. "uniform") unit-norm vector orthogonal to

$\left[\Phi(S^{(n-1)})\right]_k$. In words, the MMSE noisy update consists of applying the MMSE update (10) to all the signatures one at a time, and then adding a random bounded independent noise to each signature.

We can now prove that fixing a sufficiently small noise upper bound in the noisy iteration, $S^{(n)}$ can be made to converge to an arbitrary small neighborhood of the optimal set with probability one regardless of the initial configuration.

Theorem 1 *Given any $\delta > 0$ there exists $\theta_{max} > 0$ such that for any initial condition $S^{(0)}$ the MMSE noisy iteration defined by (34) with $\theta_{max}^{(n)} = \theta_{max}$ for all n, satisfies*

$$\limsup \mathrm{TSC}(S^{(n)}) \leq_{\text{a.s.}} \tau + \delta \tag{35}$$

Proof : We present a sketch. For details see [1]. W.l.o.g. assume δ is small enough so that if $S \in F_\Phi$ and $\mathrm{TSC}(S) \leq \tau + \delta$ then $\mathrm{TSC}(S) = \tau$. This can be done because, by lemma 2(c) the set T_F has a finite number of elements. Define the set $V \triangleq \{S \in \mathcal{S} : \mathrm{TSC}(S) \geq \tau + \delta\}$. Let

$$\theta_{max} = \min\{d(S, S') : S \in V \text{ , } S' \in \Phi(\overline{V^c})\}$$

where $\overline{V^c}$ is the closure of the complement of V. Note that θ_{max} is well-defined: $d(\cdot,\cdot)$ is a continuous function, V is a compact set, $\overline{V^c}$ is compact and thus $\Phi(\overline{V^c})$ is compact because $\Phi(\cdot)$ is continuous.

We claim $\theta_{max} > 0$ (for details see [1]).

Because of our choice of θ_{max}, if $S^{(n)} \in V^c$ then $S^{(n+1)} \in V^c$ and thus $S^{(n+k)} \in V^c$ for all $k \geq 0$.

For each $S \in \mathcal{S}$ define

$$\beta(S) \triangleq \min\{\mathrm{TSC}(S') : S' \in B[S, \theta_{max}]\}$$

Note that $\beta(S)$ is well-defined because TSC is continuous and $B[S, \theta_{max}]$ is compact. Also $\beta(S)$ is a continuous function of S because $\mathrm{TSC}(\cdot)$ is continuous and the set $B[S, \theta_{max}]$ depends continuously on S. Now define

$$\gamma \triangleq \min\{\mathrm{TSC}(S) - \beta(S) : S \in V\}$$

which is well-defined because $(\mathrm{TSC} - \beta)(\cdot)$ is continuous and $B[S, \theta_{max}]$ is compact. We claim $\gamma > 0$ ([1]).

We will write $\mathbb{P}(\cdot)$ for probabilities. For $S \in \mathcal{S}$ define

$$P(S) = \mathbb{P}\left(\mathrm{TSC}(h(\Phi(S), R, \theta)) \leq \max\left\{\mathrm{TSC}(S) - \frac{\gamma}{2}, \tau + \delta\right\}\right)$$

where $r_k, \theta_k,\ k \in \{1, \ldots, K\}$ are independent random variables, θ_k is uniform $(0, \theta_{max})$ and r_k is a random (i.e. "uniform") unit-norm vector orthogonal to $\left[\Phi(S)\right]_k$. Note that $P(S)$ is a continuous function of S because $\mathrm{TSC}(\cdot)$ is continuous and the probability distributions involved are continuous. Let

$$p \triangleq \min_{S \in V} P(S)$$

We claim $p > 0$ ([1]).

Define $M \triangleq \left(\sum_{k=1}^{K} p_k + \sum_{j=1}^{N} w_j\right)^2$. Note that $\forall S \in \mathcal{S}$, $\mathrm{TSC}(S) \leq M$. Let $L \triangleq \lceil \frac{2(M-\tau-\delta)}{\gamma} \rceil$. Write $x_n = \mathbb{P}\left(\mathrm{TSC}(S^{(nL)}) \leq \tau + \delta\right)$. It is straightforward to see

$$x_{n+1} \geq x_n + \mathbb{P}\left(\mathrm{TSC}(S^{((n+1)L)}) \leq \tau + \delta \mid \mathrm{TSC}(S^{(nL)}) > \tau + \delta\right)(1 - x_n)$$

Now

$$\mathbb{P}\left(\mathrm{TSC}(S^{((n+1)L)}) \leq \tau + \delta \mid \mathrm{TSC}(S^{(nL)}) > \tau + \delta\right) \geq p^L$$

and therefore $x_n \geq 1 - \left(1 - p^L\right)^n$. This implies that with probability 1 for some finite n_0, $S^{(n_0)} \in V^c$ and the theorem statement follows.

□

The next theorem shows that if $\theta_{max}^{(n)}$ is chosen suitably with $\theta_{max}^{(n)} \to 0$ as $n \to 0$, then $S^{(n)}$ approaches the optimal set Ω as $n \to \infty$ with probability 1.

Theorem 2 *There exists a sequence* $\theta_{max}^{(n)}$ *such that for any initial condition* $S^{(0)}$ *the MMSE noisy iteration defined by (34) satisfies*

$$\limsup \mathrm{TSC}(S^{(n)}) \to_{\text{a.s.}} \tau \tag{36}$$

Proof : Take a decreasing sequence δ_m with $lim_{m\to\infty}\delta_m = 0$, and take any $q \in (0,1)$. By the proof of theorem 1 we can find $\hat{\theta}_m$ such that, if we fix some m, the noisy MMSE update (34) with $\theta_{max}^{(n)} = \hat{\theta}_m$ satisfies $\mathbb{P}\left(\mathrm{TSC}(S^{(n)}) \leq \tau + \delta_m\right) \to 1$ as $n \to \infty$ uniformly in the initial condition $S^{(0)}$. Thus there exists l_m such that for all $S^{(0)}$ and all $n \geq l_m$, $\mathbb{P}\left(\mathrm{TSC}(S^{(n)}) \leq \tau + \delta_m\right) > q$. Let $L_m = \sum_{i=1}^{m} l_i$. It follows that if we choose $\theta_{max}^{(n)} = \hat{\theta}_m$ for all $n = (1+L_{m-1}), \ldots, L_m$, we obtain that for all $z \geq 0$ it holds $\mathbb{P}\left(\mathrm{TSC}(S^{(L_{m+z})}) \leq \tau + \delta_m\right) > 1 - (1-q)^z$. This implies $\limsup \mathrm{TSC}(S^{(n)}) \leq_{\text{a.s.}} \tau + \delta_m$ for all m. Making $m \to \infty$ we get the desired result.

5 Conclusion

The MMSE update is known to decrease the total square correlation. However, it does not guarantee convergence to the minimum TSC. We have observed that the TSC has no local minima (other than the global), and therefore the fixed configurations of the MMSE update that are not optimal are unstable. Using this fact we have proved that a modified noisy version of the MMSE update converges with probability one to the optimal.

References

[1] P. Anigstein and V. Anantharam, "Ensuring Convergence of the MMSE Iterarion for Interference Avoidance to the Global Optimum," in preparation.

[2] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*, Academic Press, 1979.

[3] C. Rose, S. Ulukus, and R.D. Yates, "Interference Avoidance for Wireless Systems," In *Vehicular Technology Conference*, May 2000.

[4] C. Rose, "Interference Avoidance, Sum Capacity and Convergence via Class Warfare," submitted to *IEEE Transactions on Information Theory.* Available at http://www.winlab.rutgers.edu/~crose.

[5] S. Ulukus and R.D. Yates, "Iterative Construction of Optimal Sequence Sets to Maximize the Capacity of CDMA Systems," submitted to *IEEE Transactions on Information Theory.* Available at http://www.winlab.rutgers.edu/~ryates.

[6] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.

[7] P. Viswanath and V. Anantharam, "Optimal Sequences and Sum Capacity of Synchronous CDMA Systems," *IEEE Transactions on Information Theory*, Vol. 45(6), pp. 1984-1991, Sept. 1999.

[8] P. Viswanath and V. Anantharam, "Total Capacity of Multiaccess Vector Channels," UC Berkeley Electronics Research Laboratory, Memorandum No. UCB/ERL M99/47, May 1999.

[9] P. Viswanath, "Capacity of Vector Multiple Access Channels," PhD. Thesis, UC Berkeley, July 2000.

Optimum Signature Sequence Sets for Asynchronous CDMA Systems

Sennur Ulukus
AT&T Labs–Research
ulukus@research.att.com

Roy D. Yates
WINLAB, Rutgers University
ryates@winlab.rutgers.edu

Abstract

In this paper, we characterize the *user capacity*, i.e., the maximum number of supportable users at a common SIR target level for a fixed processing gain, of a single cell symbol *asynchronous* CDMA system. Based on the delay profile of the users, we identify a class of optimum signature sequences that achieve a lower bound on the total squared asynchronous correlation (TSAC) among the users. When the users' signatures achieve this lower bound, the user capacity of a single-cell asynchronous CDMA system becomes the same as that of a single-cell synchronous CDMA system; that is, there is no loss in user capacity due to asynchrony. Further, when the optimum signature sequences are used, the users' received powers are all equal and the M-shot MMSE receiver filters turn out to be scaled matched filters. That is, the maximum user capacity is achieved by observing only one symbol interval of the received signal and using single-user matched filters in that interval. It is a significant open question whether these optimal sequence sets exist for all delay profiles. However, we present iterative and distributed signature adaptation algorithms, which, when executed sequentially by all of the users, appear to converge to these optimum signature sequences.

1 Introduction

Code Division Multiple Access (CDMA) systems are interference limited. Much research has been conducted in the area of multiuser detection [1] to develop techniques to suppress/cancel interference for a fixed (arbitrarily chosen) set of user signatures. Recently, there has been an interest to understand the influence of multiuser detection schemes on the overall system capacity in single-cell CDMA systems. This recent literature can be divided into two general categories: those that assume random signature sequences [2–4], and those that solve for the optimum signature sequences [5–7].

In [2, 3], a random signature sequence analysis is applied to a large system where both the number of users, K, and the processing gain, N, go to infinity, but their ratio $\alpha = K/N$ is fixed and finite. References [2, 3] showed that for this large system with random signature sequences, the signal to inteference ratio (SIR) of all users converged in probability to deterministic quantities, and calculated the *user capacity* of the system. Among all linear receiver filters, much attention is paid to the minimum mean squared error (MMSE) receiver, since it is the optimum linear filter in the sense of maximizing the SIR [8]. Reference [2] showed that the user capacity of the single-cell synchronous CDMA system with MMSE receivers is larger than that with matched filter receivers in the amount of exactly 1 user per dimension. Subsequently, [5] showed that if one can choose the signature sequences of the users carefully, then this gap between the MMSE and the matched filter receivers can be closed. The user capacity of a single-cell synchronous CDMA system using optimum signature sequences is the same for the

MMSE and matched filter receivers. Further, this capacity equals the asymptotic (i.e., large system) user capacity with random signature sequences using MMSE receivers. The signature sequences that maximize the user capacity are Welch bound equality (WBE) sequences [9–11]. In an earlier work [7], WBE sequences were shown to maximize the information theoretic sum capacity of a single-cell synchronous CDMA system with equal received powers. More recently, a generalized version of this problem where users have arbitrary (unequal) received powers was solved in [6].

An iterative and distributed signature sequence adaptation algorithm that converges to a set of optimum signature sequences was given in [12–14]. The algorithm exploits the property of the optimum signature sequences that they minimize the sum of the squares of the cross correlations among the users, called total squared correlation, TSC, in [12–14]. The algorithm converges to a set of WBE sequences if $K > N$, or a a set of orthogonal sequences if $K \leq N$. The algorithm, which was named the *MMSE update* algorithm, was based on the idea that the TSC–minimizing, hence optimum, signature sequences may be obtained if at each iteration one user updates its signature sequence to decrease the TSC of the whole set. In the proposed algorithm, each user updates its signature sequence to be the normalized MMSE receiver filter for that user when the signature sequences of all other users are fixed, hence the name MMSE update.

Recently, [4] generalized the analysis in [2] to a single-cell *asynchronous* CDMA system. Reference [4] showed that under matched filter reception, the user capacity of a single-cell asynchronous CDMA system is the same as the corresponding synchronous system. However, when MMSE reception is employed, there is a loss in user capacity compared to the user capacity in a synchronous system. This gap was shown to diminish as the observation window length is increased.

In this paper, we investigate the user capacity of an *asynchronous* single-cell CDMA system. Based on the users' delay profile, we identify a class of optimum signature sequences that achieve a lower bound on the total squared asynchronous correlation (TSAC) among the users. When the users' signatures achieve this lower bound, the user capacity of a single-cell asynchronous CDMA system is the same as that of a single-cell synchronous CDMA system, that is, there is no loss in user capacity due to asynchrony. Further, when the optimum signature sequences are used, the M-shot MMSE receiver filters turn out to be scaled matched filters; that is, the maximum user capacity is achieved by observing only one symbol interval of the received signal and using single-user matched filters in that interval. Although it is an open question whether these optimal sequence sets exist for all delay profiles, we propose iterative signature sequence adaptation algorithms which we observe experimentally to converge to optimum signature sequence sets. At each step of the algorithms only one signature sequence from the set is replaced in a way not to increase the TSAC.

2 Problem Statement and Derivations

We consider a single-cell symbol-asynchronous (but chip-synchronous) CDMA system with K users and processing gain of N. The received signal in the nth symbol interval of user k is given as

$$\mathbf{r}_k(n) = \sqrt{p_k}\, b_k(n)\mathbf{s}_k + \sum_{l\neq k} \sqrt{p_l}\left(b_l(n) T_R^{d_{kl}}\mathbf{s}_l + b_l(n+1) T_L^{d_{kl}}\mathbf{s}_l\right) \qquad (1)$$

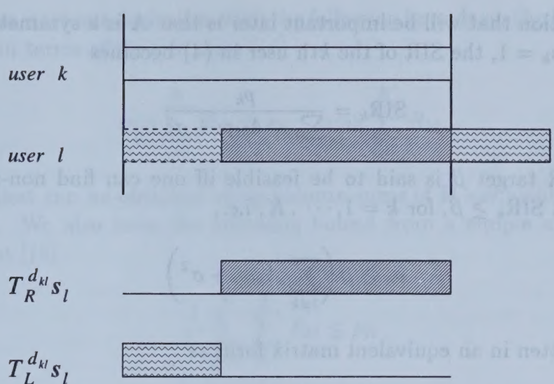


Figure 1: Asynchronous interference calculation.

where p_k, $b_k(n)$ and $\mathbf{s}_k$ are the received power, nth transmitted symbol and signature sequence of user k, respectively. The signature sequences of all users are of unit energy, i.e., $\mathbf{s}_k^\top \mathbf{s}_k = 1$, for all k. For users k and l, d_{kl} represents the relative time delay of user l with respect to the time delay of user k, that is, $d_{kl} = d_l - d_k$, where d_k and d_l are the time delays of users k and l, respectively. Symbols T_R^d and T_L^d denote the operations of shifting, to right and left, respectively, of a vector by d and $N-d$ chips (components). For both operators, the vacated positions in the vector are filled with zeros. That is, for for $d \geq 0$, we use 0^d to denote d consecutive zeros and we define

$$T_R^d \mathbf{x} = [0^d, x_1, \cdots, x_{N-d}] \tag{2}$$

and

$$T_L^d \mathbf{x} = [x_{N-d}, \cdots, x_N, 0^{N-d}] \tag{3}$$

We will use one-shot matched filters as the receivers. In general, for an arbitrary set of signature sequences, single user receivers (matched filters) are suboptimum in multiuser systems, and one-shot detectors are suboptimum in asynchronous systems. Therefore, the one-shot matched filter detector we assume is clearly suboptimum for arbitrarily chosen signature sequences. However, as we will show later, when the optimum signature sequences are used, the optimum linear receivers, i.e., the MMSE receivers, turn out to be scaled matched filters, and there is no loss in capacity due to one-shot filtering.

The decision statistics for the kth user in the nth symbol interval is $y_k(n) = \mathbf{s}_k^\top \mathbf{r}_k(n)$, where we do assume that the matched filter receiver of each user is perfectly aligned with the symbol interval of the user. The SIR of the kth user is then given by

$$\mathrm{SIR}_k = \frac{p_k(\mathbf{s}_k^\top \mathbf{s}_k)^2}{\sum_{l\neq k} p_l \left\{ (\mathbf{s}_k^\top T_R^{d_{kl}} \mathbf{s}_l)^2 + (\mathbf{s}_k^\top T_L^{d_{kl}} \mathbf{s}_l)^2 \right\} + \sigma^2(\mathbf{s}_k^\top \mathbf{s}_k)} \tag{4}$$

We define the $K \times K$ matrix $\mathbf{A}$ with the following entries

$$A_{kl} = \begin{cases} (\mathbf{s}_k^\top T_R^{d_{kl}} \mathbf{s}_l)^2 + (\mathbf{s}_k^\top T_L^{d_{kl}} \mathbf{s}_l)^2 & k \neq l \\ 0 & k = l \end{cases} \tag{5}$$

A simple observation that will be important later is that $\mathbf{A}$ is a symmetric non-negative matrix. Since $\mathbf{s}_k^\mathsf{T}\mathbf{s}_k = 1$, the SIR of the kth user in (4) becomes

$$\mathrm{SIR}_k = \frac{p_k}{\sum_{l \neq k} A_{kl} p_l + \sigma^2} \tag{6}$$

The common SIR target β is said to be feasible iff one can find non-negative powers $\{p_k\}_{k=1}^K$ such that $\mathrm{SIR}_k \geq \beta$, for $k = 1, \cdots, K$, i.e.,

$$p_k \geq \beta \left(\sum_{l \neq k} A_{kl} p_l + \sigma^2 \right) \tag{7}$$

which can be written in an equivalent matrix form as

$$\mathbf{p} \geq \beta \left(\mathbf{A}\mathbf{p} + \sigma^2 \mathbf{1} \right) \tag{8}$$

where $\mathbf{1}$ is the vector of all ones. It is well known that if the SIR target β is feasible, then the optimum power vector, i.e., the componentwise smallest feasible power vector, is found by solving (8) with equality:

$$\mathbf{p} = \beta \mathbf{A}\mathbf{p} + \sigma^2 \beta \mathbf{1} \tag{9}$$

That is, if the common SIR target β is feasible, the optimum power vector is given as $\mathbf{p}^* = \sigma^2 \beta (\mathbf{I} - \beta \mathbf{A})^{-1} \mathbf{1}$. The power control problem is feasible iff [15]

$$\beta < \frac{1}{\rho_A} \tag{10}$$

where ρ_A is the largest (also called the Perron-Frobenius) eigenvalue of matrix $\mathbf{A}$. We define the matrix $\mathbf{R} = \mathbf{A} + \mathbf{I}$ so that $\mathbf{R}_{kk} = (\mathbf{s}_k^\mathsf{T}\mathbf{s}_k)^2 = 1$ and $\mathbf{R}$ represents the squared asynchronous cross correlations of the signatures. The Perron-Frobenius eigenvalue of $\mathbf{R}$ satisfies $\rho_R = \rho_A + 1$, and the feasibility condition in (10) can also be expressed as

$$\beta < \frac{1}{\rho_R - 1} \tag{11}$$

That is, for a single cell CDMA system, the range of common achievable SIR values are determined only by the Perron-Frobenius eigenvalue of the squared asynchronous cross correlation matrix $\mathbf{R}$ which depends only on the signature sequences of the users and their relative time delays. For a given signature sequence set $\{\mathbf{s}_k\}_{k=1}^K$ and a set of time delays $\{d_k\}_{k=1}^K$, the supremum of common achievable SIR targets equals $1/(\rho_R - 1)$ Our aim is to choose the signature sequences of the users, for any given set of time delays, such that the common achievable SIR is maximized. Therefore, we seek the signature sequence set that maximizes $1/(\rho_R - 1)$, or, equivalently, minimizes ρ_R.

We note that it is hard to characterize the dependence of ρ_R on individual signature sequences. If this was not the case, one could devise an algorithm to update the signature sequences of the users in the direction that decreases ρ_R. Instead, our approach is to tie the Perron-Frobenius eigenvalue of $\mathbf{R}$, ρ_R, to another parameter which can be related to the signature sequences in a more direct way. By this approach we will be able to characterize the optimum signature sequences in a closed form expression in addition to being able to devise an iterative and distributed signature sequence update algorithm that will construct progressively better signature sequence sets.

To this end, we start our derivation with the following bounds on the Perron-Frobenius eigenvalue of $\mathbf{R}$ in terms of its row-sums [15]

$$\min_k \sum_{l=1}^{K} R_{kl} \leq \rho_R \leq \max_k \sum_{l=1}^{K} R_{kl} \tag{12}$$

Similar bounds that can be obtained using column-sums of $\mathbf{R}$ are identical to (12) since $\mathbf{R}$ is symmetric. We also have the following bound from a simple application of the Rayleigh quotient [16]

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K} R_{kl} \leq \rho_R \tag{13}$$

which is equivalent to $(\mathbf{1}^\top \mathbf{R}\mathbf{1})/(\mathbf{1}^\top \mathbf{1}) \leq \rho_R$. Combining (12) and (13) and the fact that the minimum row-sum lower bounds the average of the row-sums yields

$$\min_k \sum_{l=1}^{K} R_{kl} \leq \frac{1}{K}\sum_{k=1}^{K}\sum_{l=1}^{K} R_{kl} \leq \rho_R \leq \max_k \sum_{l=1}^{K} R_{kl} \tag{14}$$

We define the total squared asynchronous correlation (TSAC) as

$$\mathrm{TSAC} = \sum_{k=1}^{K}\sum_{l=1}^{K} R_{kl} \tag{15}$$

Since we want to minimize ρ_R, and since ρ_R is lower bounded by TSAC/K, it is reasonable to try to minimize the TSAC over the space of all possible signature sequences. It is not clear, however, that ρ_R decreases as TSAC decreases. We will show, though, that the signature sequence sets that minimize TSAC are precisely those that minimize ρ_R.

In order to motivate the solution of the asynchronous problem, we will first revisit the synchronous problem which has been recently solved in [5].

3 The Synchronous Problem Revisited

In the synchronous case the (k, l)th component of $\mathbf{R}$ is $R_{kl} = (\mathbf{s}_k^\top \mathbf{s}_l)^2$. The following two theorems guarantee that the signature sequences that minimize the TSC (equivalent of TSAC in synchronous case) are those that minimize ρ_R.

Theorem 1 (Welch [9], Massey [10], Massey-Mittelholzer [11])

$$TSC = \sum_{k=1}^{K}\sum_{l=1}^{K} (\mathbf{s}_k^\top \mathbf{s}_l)^2 \geq \frac{K^2}{N} \tag{16}$$

Theorem 2 (Massey-Mittelholzer [11]) *(The Uniformly-Good Property) If the sequences $\{\mathbf{s}_k\}_{k=1}^{K}$ are such that the equality holds in (16) then*

$$\sum_{l=1}^{K} (\mathbf{s}_k^\top \mathbf{s}_l)^2 = \frac{K}{N} \qquad k = 1, \cdots, K \tag{17}$$

Theorem 1, combined with (14) and $R_{kl} = (\mathbf{s}_k^\top \mathbf{s}_l)^2$ for the synchronous case, says that

$$\rho_R \geq \frac{K}{N} \tag{18}$$

Since our aim is to minimize ρ_R, the best we can do is to choose signature sequences so as to achieve (18) with equality. Theorem 2 says that when the signature sequences are chosen such that the TSC is minimized, i.e., the bound on the TSC is satisfied by equality, then all of the row-sums are equal, and they are all equal to K/N. Since the row-sums sandwich ρ_R, (18) is also satisfied with equality, and the lowest possible ρ_R is obtained: $\rho_R = K/N$. Therefore, using (11), in the synchronous case, the bound on the common achievable SIR target is

$$\beta < \frac{1}{K/N - 1} \tag{19}$$

which is equivalent to the user capacity expression

$$\frac{K}{N} < 1 + \frac{1}{\beta} \tag{20}$$

given in [5] where this problem was originally solved.

Theorems 1 and 2 apply to $K > N$ case. When $K \leq N$, the bound in Theorem 1 is loose, the K^2/N bound cannot be achieved, and Theorem 2 loses its applicability. When $K \leq N$, the equivalent of (16) in Theorem 1 is

$$\text{TSC} \geq K \tag{21}$$

In this case, the equivalent of Theorem 2 is the following: if the sequences $\{\mathbf{s}_k\}_{k=1}^K$ are such that the equality holds in (21) then $\sum_l (\mathbf{s}_k^\top \mathbf{s}_l)^2 = 1$ for $k = 1, \cdots, K$. That is, all of the row-sums of $\mathbf{R}$ are equal to 1, and therefore, $\rho_R = 1$. The implication of this result, from (11), is that any (arbitrarily large) common SIR target, β, is feasible with sufficient transmitter power.

4 The Asynchronous Problem

In this section, we derive asynchronous versions of Theorems 1 and 2. The following two theorems guarantee that the signature sequences that minimize the TSAC are those that minimize ρ_R.

Theorem 3 *For any delay profile, the total squared asynchronous correlation satisfies*

$$TSAC = \sum_{k=1}^{K} \sum_{l=1}^{K} R_{kl} \geq \frac{K^2}{N} \tag{22}$$

Theorem 4 *(The Asynchronous Uniformly-Good Property) If the sequences* $\{\mathbf{s}_k\}_{k=1}^K$ *are such that the equality holds in (22) then*

$$\sum_{l=1}^{K} R_{kl} = \frac{K}{N} \qquad k = 1, \cdots, K \tag{23}$$

We will continue our derivation similar to the synchronous case. For this asynchronous CDMA system, Theorem 3 combined with (14) says that

$$\rho_R \geq \frac{K}{N} \tag{24}$$

Similar to the synchronous case, our aim is to minimize ρ_R, and we cannot do better than to choose signature sequences that achieve (24) with equality. Theorem 4 says that when the signature sequences are chosen such that the TSAC is minimized, i.e., the bound on the TSAC is satisfied by equality, then all of the row-sums are equal, and they are all equal to K/N. Since the row-sums sandwich ρ_R, (24) is also satisfied with equality, and the lowest possible ρ_R is obtained: $\rho_R = K/N$. Therefore, using (11), the bound on the common achievable SIR target, in this asynchronous case, is

$$\beta < \frac{1}{K/N - 1} \tag{25}$$

which is the same as the bound (19) found in the synchronous case.

Similar to the synchronous case, Theorems 3 and 4 apply to the $K > N$ case. When $K \leq N$, the bound in Theorem 3 is loose, the K^2/N bound cannot be achieved, and Theorem 4 loses its applicability. When $K \leq N$, the equivalent of (22) in Theorem 3 is

$$\text{TSAC} \geq K \tag{26}$$

The bound is achieved with equality when A_{kl} satisfies $A_{kl} = 0$ for all k and l. In this case, the equivalent of Theorem 4 is the following: if the sequences $\{\mathbf{s}_k\}_{k=1}^K$ are such that the equality holds in (26) then $\sum_l R_{kl} = 1$ for $k = 1, \cdots, K$. That is, all of the row-sums of $\mathbf{R}$ are equal to 1, and therefore, $\rho_R = 1$. The implication of this result, from (11), as in the synchronous case, is that any (arbitrarily large) common SIR target, β, is feasible.

We will next show that the optimum received powers of the users are equal. First, we set the received power of each user k to be

$$p_k = p = \frac{\sigma^2}{1 + 1/\beta - K/N} \qquad k = 1, \cdots, K \tag{27}$$

and then we show that this selection guarantees that all users have SIRs equal to β when the signature sequences of the users are chosen to satisfy (22) with equality in Theorem 3. Note that $p > 0$ as long as K, N and β satisfy the user capacity inequality (25). From (6), the SIR of the kth user is

$$\text{SIR}_k = \frac{p}{p(\sum_{l \neq k} A_{kl}) + \sigma^2} = \frac{p}{p(\sum_l R_{kl} - 1) + \sigma^2} \tag{28}$$

If the signature sequences satisfy (22) with equality in Theorem 3, then using (23) in Theorem 4, we can write (28) as

$$\text{SIR}_k = \frac{p}{p(K/N - 1) + \sigma^2} \tag{29}$$

It is now straightforward to show that when (27) is inserted into (29), we have $\text{SIR}_k = \beta$.

As stated earlier, in general, using the matched filter as we have done in this section for the asynchronous case and in the previous section in the treatment of the synchronous case is suboptimum. However, an important observation in the original solution for the optimum signature sequences in the synchronous case in [5] was that once the optimum signature sequences were chosen, the MMSE receiver filters reduce to scaled matched filters. As stated in the following theorem, when we choose the optimum signature sequences in the asynchronous case, i.e., the signature sequences that minimize the TSAC, then the M-shot MMSE receiver filters, for $M \geq 1$, corresponding to those signature sequences reduce to be scaled matched filters.

Theorem 5 *If the sequences* $\{\mathbf{s}_k\}_{k=1}^{K}$ *are such that the equality holds in (22) then the* M*-shot MMSE filters for all users reduce to corresponding scaled matched filters.*

5 TSAC Reduction: Iterative Algorithms

Following the closed-form expressions for the signature sequence sets maximizing the information theoretic sum capacity [6,7] and user capacity [5], references [12–14] introduced the iterative adaptation of signature sequences for synchronous CDMA systems. Since the optimum signature sequences minimize the TSC in the synchronous case, the algorithm presented in [12–14] was designed to decrease (more precisely, not to increase) the TSC at each iteration of the algorithm. Since we have shown in the previous section that the optimum signature sequences minimize the TSAC in the asynchronous case, we will design algorithms which decrease the TSAC at each iteration. To this end, we will first separate the terms that depend on the signature sequence of the kth user in the TSAC. Using the TSAC definition (15) and the definition of $\mathbf{R}$ as $\mathbf{R} = \mathbf{A} + \mathbf{I}$, along with the definition of $\mathbf{A}$ in (5), we can express the TSAC as

$$\text{TSAC} = (\mathbf{s}_k^{\mathsf{T}}\mathbf{s}_k)^2 + 2\mathbf{s}_k^{\mathsf{T}}\left(\sum_{l\neq k} \bar{\mathbf{s}}_{kl}\bar{\mathbf{s}}_{kl}^{\mathsf{T}} + \tilde{\mathbf{s}}_{kl}\tilde{\mathbf{s}}_{kl}^{\mathsf{T}}\right)\mathbf{s}_k + \gamma_k \tag{30}$$

where we used notation $\bar{\mathbf{s}}_{kl} = T_R^{d_{kl}}\mathbf{s}_l$ and $\tilde{\mathbf{s}}_{kl} = T_L^{d_{kl}}\mathbf{s}_l$, to represent the left and right signatures of the lth asynchronous user with respect to the kth user. We also used the fact that $\mathbf{A}$ is symmetric in deriving (30) from the definition of TSAC in (15). In (30),

$$\gamma_k = \sum_{i\neq k}\sum_{j\neq k} R_{kl} \tag{31}$$

denotes the squared asynchronous correlation terms that do not depend on $\mathbf{s}_k$. Let us define $\mathbf{B}_k$ as

$$\mathbf{B}_k = \sum_{l\neq k}\left(\bar{\mathbf{s}}_{kl}\bar{\mathbf{s}}_{kl}^{\mathsf{T}} + \tilde{\mathbf{s}}_{kl}\tilde{\mathbf{s}}_{kl}^{\mathsf{T}}\right) \tag{32}$$

Therefore the TSAC in (30) can be written as

$$\text{TSAC} = (\mathbf{s}_k^{\mathsf{T}}\mathbf{s}_k)^2 + 2\mathbf{s}_k^{\mathsf{T}}\mathbf{B}_k\mathbf{s}_k + \gamma_k \tag{33}$$

In order to minimize the TSAC, we are looking for updates of the signature sequence of the kth user from $\mathbf{s}_k$ to some $\mathbf{c}_k$ that is guaranteed to decrease (not to increase) the TSAC. Let us denote the TSAC after the $\mathbf{s}_k \rightarrow \mathbf{c}_k$ update as $\overline{\text{TSAC}}$. Then,

$$\overline{\text{TSAC}} = (\mathbf{c}_k^{\mathsf{T}}\mathbf{c}_k)^2 + 2\mathbf{c}_k^{\mathsf{T}}\mathbf{B}_k\mathbf{c}_k + \gamma_k \tag{34}$$

Restricting the new (updated) signature sequence of the kth user to be of unit energy as well, i.e., $\mathbf{c}_k^\top \mathbf{c}_k = 1$, we note that $\overline{\text{TSAC}} \leq \text{TSAC}$ iff

$$\mathbf{c}_k^\top \mathbf{B}_k \mathbf{c}_k \leq \mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k \tag{35}$$

Although there are many possible $\mathbf{s}_k \to \mathbf{c}_k$ updates that would guarantee that (35) holds, we will propose two of them here. The two similar updates used in the synchronous CDMA context were given in [12–14] and in [17]. We call the first update the *asynchronous MMSE update* which is given as

$$\mathbf{c}_k = \frac{\left(\mathbf{B}_k + a^2 \mathbf{I}_N\right)^{-1} \mathbf{s}_k}{\left[\mathbf{s}_k^\top \left(\mathbf{B}_k + a^2 \mathbf{I}_N\right)^{-2} \mathbf{s}_k\right]^{1/2}} \tag{36}$$

and we call the second update the *asynchronous eigen update* which is defined as the normalized version of the eigenvector of $\mathbf{B}_k$ corresponding to its smallest eigenvalue. The proof that the asynchronous eigen update decreases the TSAC follows from the Rayleigh quotient applied to the matrix $\mathbf{B}_k$ [16]. The proof that the asynchronous MMSE update decreases the TSAC can be carried out in a very similar fashion to the proof that the MMSE update decreases the TSC in [12–14].

6 Further Remarks

Due to space limitations, we were unable to provide proofs of our theorems in this paper. The derivations of the bounds in the proofs of Theorems 3 and 4 describe the conditions that the signature sequences must satisfy in order for them to achieve the lower bound on the TSAC. In a synchronous CDMA system, the optimum (WBE) sequences must satisfy $\mathbf{S}\mathbf{S}^\top = (K/N)\mathbf{I}_N$, where $\mathbf{S}$ is a matrix containing the signature sequences of the users in its columns. That is, the optimum signature sequences must satisfy three sets of conditions: (1) each column of $\mathbf{S}$ must have unit length, (2) rows of $\mathbf{S}$ must be orthogonal to each other, and (3) each row of $\mathbf{S}$ must have equal length (a length of $\sqrt{K/N}$). These conditions were identified in [10,11] in re-deriving the Welch's bound, but the existence of sequences satisfying these conditions was first addressed in [5]. Similar to [10, 11] in the synchronous case, currently, in the asynchronous case, we have derived the conditions that the optimum signature sequences should satisfy, however, we have not addressed the issue of existence of such sequences. Similarly, we have proved that the optimum signature sequences should achieve a lower bound on TSAC and provided two algorithms that monotonically decrease TSAC; however we have not proved that the algorithms proposed here are guaranteed to converge to the desired lower bound starting from an arbitrary set. We can report however that, through a large number of numerical experiments with randomly generated initial signature sequence sets, we have observed that the TSAC reduction algorithms we have proposed here have always converged to signature sequences with TSAC $= K^2/N$. That is, we have observed not only the existence of such sequence sets, but also the convergence of the proposed algorithms to these sets.

References

[1] S. Verdú. *Multiuser Detection.* Cambridge University Press, 1998.

[2] D. Tse and S. V. Hanly. Linear multiuser receivers: Effective interference, effective bandwidth and user capacity. *IEEE Trans. on Information Theory*, 45(2):641–657, March 1999.

[3] S. Verdú and S. Shamai. Spectral efficiency of CDMA with random spreading. *IEEE Trans. on Information Theory*, 45(2):622–640, March 1999.

[4] Kiran and D. Tse. Effective interference and effective bandwidth of linear multiuser receivers in asynchronous CDMA systems. *IEEE Trans. on Information Theory*, 46(4):1426–1447, July 2000.

[5] P. Viswanath, V. Anantharam, and D. Tse. Optimal sequences, power control, and user capacity of sysnchronous CDMA systems with linear MMSE multiuser receivers. *IEEE Trans. on Information Theory*, 45(6):1968–1983, September 1999.

[6] P. Viswanath and V. Anantharam. Optimal sequences and sum capacity of synchronous CDMA systems. *IEEE Trans. on Information Theory*, 45(6):1984–1991, September 1999.

[7] M. Rupf and J. L . Massey. Optimum sequence multisets for synchronous code-division multiple-access channels. *IEEE Trans. on Information Theory*, 40(4):1261–1266, July 1994.

[8] U. Madhow and M. L. Honig. MMSE interference suppression for direct-sequence spread-spectrum CDMA. *IEEE Trans. on Communications*, 42(12):3178–3188, December 1994.

[9] L. R. Welch. Lower bounds on the maximum cross correlation of signals. *IEEE Trans. on Information Theory*, IT-20(3):397–399, May 1974.

[10] J. L. Massey. On Welch's bound for the correlation of a sequence set. In *IEEE ISIT*, page 385, 1991.

[11] J. L . Massey and T. Mittelholzer. Welch's bound and sequence sets for code-division multiple-access systems. *Sequences II: Methods in Communication, Security and Computer Science.* R. Capocelli, A. De Santis and U. Vaccaro, Eds. New York: Springer-Verlag, 1991.

[12] S. Ulukus. Power control, multiuser detection and interference avoidance in CDMA systems. Ph.D. Thesis, Dept. of Electrical and Computer Engineering, Rutgers University, NJ. July 1998. Available at http://www.research.att.com/~ulukus.

[13] S. Ulukus and R. D. Yates. Iterative signature adaptation for capacity maximization of CDMA systems. In *36th Annual Allerton Conference on Communications, Control and Computing*, September 1998.

[14] S. Ulukus and R. D. Yates. Iterative construction of optimum signature sequence sets in synchronous CDMA systems. IEEE Trans. on Information Theory, Submitted Sept. 1999; Revised Sept. 2000. Available at http://www.research.att.com/~ulukus.

[15] E. Seneta. *Non-negative Matrices and Markov Chains.* Springer Verlag, 1981.

[16] G. Strang. *Linear Algebra and Its Applications.* Saunders College Publishing, 1988.

[17] C. Rose, S. Ulukus, and R. D. Yates. Interference avoidance for wireless systems. In *IEEE VTC 2000*, May 2000.

Randomized Power Control Strategies for Optimization of Multiple Access Radio Systems *

Guillermo del Angel, Terrence L. Fine
Cornell University.
School of Electrical and Computer Engineering.
Ithaca, NY 14853
{delangel,tlfine}@ee.cornell.edu

Abstract

We introduce a dynamic control scheme to optimize access to a Slotted Aloha random access protocol in a mobile radio system. In our model, a population of users attempts to transmit packets to a central receiver/controller during successive time slots. If a packet is received with a signal-to-interference ratio which exceeds a certain threshold, the packet is captured and is received successfully. With this model, an optimal randomized power and retransmission control strategy is derived, in which the transmitters choose a random power level for their packet transmission according to a centrally derived probability distribution. The control strategy is determined at each time slot based on the feedback that the receiver gets from the transmitters. The optimal power levels and probabilities are derived, and it is shown that, given the capture model, the optimal probability distribution for the power levels is purely discrete. Theoretical bounds for the achievable throughput are given and the proposed system performance is explored through numerical simulations.

1 Introduction

We consider in this work a dynamic control scheme to optimize access to a radio system. Specifically, we consider a situation in which several users attempt to transmit information to a central receiver using a multiple access protocol such as Slotted Aloha with retransmission control, to be described later.

In a mobile radio system, the received signal power will vary among the transmitters due to differences in spatial location and channel conditions. These differences may create what is referred to as the Capture effect, in which a packet may be successfully received even in the presence of other transmitters if its corresponding signal is sufficiently strong compared to others. These variations might actually improve the performance of a random access protocol such as Slotted Aloha [5, 8].

To exploit received power variability, randomized strategies have been proposed, in which transmitters explicitly vary their transmission power in order to induce capture at the receiver [7, 4]. Our model extends these results in order to consider an optimal dynamic transmit power and retransmission control policy, considering additional elements such as power level construction and throughput bounds not considered before.

---

*Prepared with partial support from NSF grant NCR-9725251 and Air Force Grant F30602-00-2-0558. G. del Angel was also supported in part by CONACYT (Mexico) doctoral grant #116047

This work is outlined as follows. In Section 2 we give a mathematical description of the system model. Section 3 develops optimal strategies for power and retransmission control. Section 4 analyzes and proves results dealing with power level construction. In Section 5 we present bounds on the achievable system throughput. In Section 6 we present simulation results to illustrate the proposed system performance, and finally conclusions and plans for future work are given in Section 7.

2 Mathematical Model

We consider a slotted Aloha system, in which all time slots have the same duration and where, for simplicity, we assume perfect synchronization between all transmitters and a central receiver/controller.

The total user population may be finite (say, of M users), or infinite. A user may be in one of two states at the beginning of each time slot, either idle or backlogged. If a user is idle, it generates a new packet with probability q_a for the case of a finite user population, or according to a Poisson process with arrival rate λ for an infinite user population. The new packet generation is assumed to be of constant rate and independent among time slots. Let A_t denote the number of new transmitters at slot t and let n_t denote the number of backlogged users at the beginning of slot t. For a finite user case, we can write $\Pr(A_t = k|n_t) = b(M - n_t, k, q_a), k \leq M - n_t$ with $b(n, k, p) = \binom{n}{k} p^k (1-p)^{n-k}$. For an infinite user population, we can write $\Pr(A_t = k) = e^{-\lambda}\lambda^k/k!$ In the case of a finite user population, it is assumed that users do not buffer packets as to simplify analysis.

If an idle user's transmission collides with other users' and the transmission is not received successfully at the central receiver, the user becomes backlogged. Backlogged users at the beginning of slot t attempt to retransmit their packet with probability q_t. We assume that retransmission attempts are independent among backlogged users and that q_t is the same among them. Let R_t denote the number of retransmissions at slot t. Then, clearly we can write $\Pr(R_t = k|n_t) = b(n_t, k, q_t), k \leq n_t$

At the end of the time slot, a feedback message $\hat{f}_t$ is generated giving some information to the system about the outcome of the transmissions. We will assume that the feedback process only depends on the number of transmitters at the current time slot $f_t = A_t + R_t$ and on the successful reception of a packet at the receiver, indicated by I_{C_t}, and thus is completely characterized by the conditional probability $\Pr(\hat{f}_t|f_t, I_{C_t})$ and is independent of n_t given f_t. For example, it may be possible for the receiver to infer an estimate of f_t from other information, such as power distribution throughout a cell learned from a sensor array[1].

In Section 6, we study this model numerically in the following two cases:

- *Perfect Feedback*: The number of transmitters f_t is known, so $\hat{f}_t = f_t$.
- *Ternary Feedback*: It is only known whether there was silence on the channel or there was any transmission, and in the last case, if the packet reception was successful or not.

As noted before, a packet is captured if it is received with a power significantly higher than the rest. To put it precisely, let P_t^k denote the received signal power at time slot t for transmitter k. Let N be the mean background noise power (assumed to be constant) and let f_t be the number of transmitters at slot t. We consider then that a packet from user k is captured iff

$$P_t^k \geq \alpha \left(\sum_{j=1, j \neq k}^{f_t} P_t^j + N \right) \tag{1}$$

The threshold constant α is system-dependent and in general would depend on physical layer characteristics such as modulation, coding, etc. We will assume that $\alpha \geq 1$. Thus, Eq. 1 is equivalent to stating that capture occurs iff the signal to interference ratio (SIR) for a transmitter exceeds α.

Note that the capture model of Eq. 1 is a generalization of classical Aloha, as by setting $N = 0$ and $\alpha \to \infty$ we get the classical assumption of either perfect packet reception (if $n = 1$) or packet destruction for $n > 1$.

In our approach, if a user attempts a packet transmission at time slot t, it transmits it with a random power level determined according to a certain probability distribution, independently of other transmitters. An important assumption is that the transmitters can exercise sufficiently good closed-loop power control, so that all of them can scale their transmitted powers as to make the received power at the central station lie within the same range for all users, which we can take without loss of generality to be the interval $[0, 1]$. Thus, power variations due to distance and shadowing are eliminated, and fading is neglected. Hence, the signal powers at the central receiver are independent, identically distributed random variables with a common distribution function $F_t(x) = \Pr(P_t^k \leq x)$. This setup ensures fairness among users, since without proper scaling a packet from a transmitter very close to the receiver would have a much higher probability of being captured.

As in the case of classical Slotted Aloha, the system dynamics can be described by the number of backlogged users at each time slot. It can be seen from the previous assumptions that n_t evolves as $n_{t+1} = n_t + A_t - I_{C_t}$ where C_t denotes a capture event for slot t and I_A is the indicator function for event A. In general, the value for n_t will be unknown to the central receiver. In order to make a stochastic model for n_t, note that n_{t+1} is conditionally independent of all $n_{t'}$ and $\hat{f}_{t'}$ for $t' < t$ given n_t.

We will show in Section 3 that, under the capture model of Equation 1, there exists an optimal distribution which is purely discrete, with finite support, and independent of the number of transmitters. It is then sufficient to consider a fixed finite number of power levels, depending only on α and N without losing optimality. We can characterize this distribution then by a vector $\mathbf{L}$ of m levels (the structure of which will be discussed later on) and their corresponding probability, determined by a probability vector ϕ_t with components $\phi_t(i) = \Pr(P_t^k = L_i)$.

3 Optimal Control Strategies

At the beginning of slot t, the central receiver broadcasts the retransmission probability q_t as well as the power level probabilities ϕ_t. The transmission time for these messages is neglected and we assume that these transmissions are received perfectly by all users.

This control mechanism is more complex than classical retransmission control of [2, 8] where only q_t is varied, but it fully exploits the capture phenomenon.

We are interested in optimal values for the control parameters (ϕ_t, q_t), such that some measure of the system performance is maximized. We will consider here a simple greedy optimization criterion, which is to maximize the throughput S of the system given all the past information available to the controller. Since a packet is successfully received iff there is capture, we can express $S = \Pr(\mathcal{C}_t)$.

Let $\mathbf{h}_t$ denote all the data available to the receiver up to the beginning of slot t, so $\mathbf{h}_t = \{\hat{f}_1, \ldots, \hat{f}_{t-1}, I_{\mathcal{C}_1}, \ldots I_{\mathcal{C}_{t-1}}\}$ Thus, the proposed control law takes the form

$$(\phi_t^*, q_t^*) = \operatorname*{argmax}_{\phi_t \in \Delta_m, q_t \in [0,1]} \Pr(\mathcal{C}_t | \mathbf{h}_t; \phi_t, q_t) \tag{2}$$

where Δ_m is the unit simplex of dimension m. Define the (possibly infinite) vector $\boldsymbol{\pi}_t$ with components $\pi_t(i) = \Pr(n_t = i | \mathbf{h}_t)$. The vector $\boldsymbol{\pi}_t$ itself forms a Markov process with uncountable state space, which is known to the receiver, as opposed to n_t which is unknown, and it summarizes the uncertainty about n_t given all available data. It is a sufficient statistic for the control process of Eq. 2.

Based on the model described in Section 2, we can update $\boldsymbol{\pi}_t$ recursively when $\hat{f}_t$ and $I_{\mathcal{C}_t}$ are observed. First, note that $\pi_{t+1}(j) = \sum_i \Pr(n_{t+1} = j | \hat{f}_t = k, n_t = i, I_{\mathcal{C}_t} = c) \Pr(\hat{f}_t = k, I_{\mathcal{C}_t} = c | n_t = i) \pi_t(i) / \Pr(\hat{f}_t = k, I_{\mathcal{C}_t} = c | \mathbf{h}_t)$. Hence,

$$\pi_{t+1}(j) = \frac{\sum_i \sum_n \Pr(\hat{f}_t = k | f_t = n, I_{\mathcal{C}_t} = c) \Pr(n_{t+1} = j, f_t = n, I_{\mathcal{C}_t} = c | n_t = i) \pi_t(i)}{\sum_j \sum_i \sum_n \Pr(\hat{f}_t = k | f_t = n, I_{\mathcal{C}_t} = c) \Pr(n_{t+1} = j, f_t = n, I_{\mathcal{C}_t} = c | n_t = i) \pi_t(i)} \tag{3}$$

The probability $\Pr(n_{t+1} = j, f_t = n, I_{\mathcal{C}_t} = c | n_t = i)$ can be found as follows, letting $C_n = \Pr(I_{\mathcal{C}_t} = 1 | f_t = n)$ (which depends on the received power distribution F_t):

$$\begin{aligned} &\Pr(n_{t+1} = j, f_t = n, I_{\mathcal{C}_t} = c | n_t = i) = \\ &\Pr(I_{\mathcal{C}_t} = c | f_t = n) \Pr(A_t = j + c - i | n_t = i) \Pr(R_t = n - j - c + i | n_t = i) = \\ &\begin{cases} 0 & j < i - 1 \\ C_n \Pr(A_t = j - i + 1 | n_t = i) \Pr(R_t = n - j + i - 1 | n_t = i) & i \le j + 1 \le i + n, c = 1 \\ (1 - C_n) \Pr(A_t = j - i | n_t = i) \Pr(R_t = n - j + i | n_t = i) & i \le j \le i + n, c = 0 \\ 0 & j > i + n \end{cases} \end{aligned} \tag{4}$$

From Eq. 1, it is clear that the capture probability only depends on f_t, α, N and the distribution for the received powers $F_t(x)$. Also, $\Pr(f_t = n | n_t = k) = \Pr(R_t + A_t = n | n_t = k) = \sum_{j=0}^{n} \Pr(R_t = j | n_t = k) \Pr(A_t = n - j | n_t = k)$. Then, $\Pr(\mathcal{C}_t | \mathbf{h}_t) = \sum_k \sum_n C_n \Pr(f_t = n | n_t = k) \pi_t(k)$.

The calculation of C_n is, however, more elaborate. From the capture criterion of Eq. 1, the probability that a packet is captured given $f_t = n$ transmitters is:

$$C_n = n \Pr\left(P_t^k \ge \alpha \left(\sum_{1 \le j \le n, j \ne k} P_t^j + N\right)\right) = n \int_{x \in [0,1]} \Gamma^n\left(\frac{x}{\alpha} - N\right) dF_t(x) \tag{5}$$

where $\Gamma^n(x)$ is the $(n-1)$-fold convolution of $F_t(x)$ and where we used the fact that all received user powers are assumed to be i.i.d. with common distribution function F_t.

Equation 5 is difficult to evaluate in general and it can have closed form expressions only for a few special cases. For the case when the distribution F_t is discrete, the evaluation simplifies significatively. For a vector of m power levels $\mathbf{L} = \{L_1, \dots, L_m\}$ with corresponding probability vector $\boldsymbol{\phi}_t$ with elements $\phi_t(i) = \Pr(P_t^k = L_i)$, define the set $\mathcal{K}_n$ as the collection of realizations of power level choices which yield capture:

$$\mathcal{K}_n(\alpha, N, \mathbf{L}) = \left\{ \mathbf{k} \in \mathbb{N}^m : \sum_i k_i = n \text{ and } \exists j : k_j = 1, L_j \geq \alpha \left(\sum_{i \neq j} k_i L_i + N \right) \right\} \tag{6}$$

Then, Eq. 5 can be expressed as [4]:

$$C_n = \Pr(\mathcal{C}_t | f_t = n; \boldsymbol{\phi}) = \sum_{\mathbf{k} \in \mathcal{K}_n} \binom{n}{k_1\ k_2\ \dots\ k_m} \prod_{i=1}^{m} \phi_t(i)^{k_i} \tag{7}$$

The number of terms in the summation in Eq. 7 is upper bounded in [4] by $\binom{n+m-1}{m-1}$. Note that, once $\mathbf{L}$ is fixed, the set $\mathcal{K}_n$ is constant, and thus need not be recomputed online if the noise power N is not time varying.

With the above formulas, then, the control algorithm takes a two-step form. For slot $t+1$:

1. Update state vector $\boldsymbol{\pi}_t$ according to Eq. 3 based on the actions and observations from slot t. Denote this operation as: $\boldsymbol{\pi}_{t+1} = U(\boldsymbol{\pi}_t, \hat{f}_t, I_{\mathcal{C}_t}, q_t, \boldsymbol{\phi}_t)$.
2. Determine control parameters as
$$(\boldsymbol{\phi}_{t+1}^*, q_{t+1}^*) = \underset{\boldsymbol{\phi}_{t+1} \in \Delta_m, q_{t+1} \in [0,1]}{\operatorname{argmax}} \sum_k \Pr(\mathcal{C}_{t+1} | n_{t+1} = k; \boldsymbol{\phi}_{t+1}, q_{t+1}) \boldsymbol{\pi}_{t+1}(k)$$

Now, we show that we do not lose performance by considering a finite number of levels $\mathbf{L}$, depending only on N and α.

4 Power Level Construction

Assume that the number of transmitters during slot t is $f_t = n$. Let $\mathbf{x} = \{x_1, \dots, x_n\} \in [0,1]$ be a sequence of i.i.d. random variables representing the received powers for n transmitters, all of them characterized by the distribution function $F_t(x)$.

Define the order statistics of $\mathbf{x}$ as $x_j^{(n)}$, s.t. $x_1^{(n)} \leq \dots \leq x_n^{(n)}$. So, for example, $x_n^{(n)} = \max_j x_j$.

Let $\mathcal{C}^n$ be the set of realizations of $\mathbf{x}$ s.t. capture occurs, with corresponding probability $C_n = \Pr(\mathcal{C}^n; F_t)$ determined by F_t according to Eq. 5:

$$\mathcal{C}^n = \left\{ \mathbf{x} \in [0,1]^n : x_n^{(n)} \geq \alpha \left(\sum_{j=1}^{n-1} x_j^{(n)} + N \right) \right\}$$

If $\mathbf{x} \in \mathcal{C}^n$ we say that $\mathbf{x}$ is a **capture event**.

Theorem 1. Let $\alpha \geq 1$, $n \geq 1$ and $N > 0$ be given. Given any probability measure for $x \in [0,1]$ defined by a distribution function F_t, there exists another probability measure defined by a probability vector $\boldsymbol{\phi}_t$ s.t.:

1. $\Pr(\mathcal{C}^n; \phi_t) \geq \Pr(\mathcal{C}^n; F_t)$
2. ϕ_t has a finite support.

Proof. Let $\mathcal{L}_n$ be a class of sets with the following properties: A set $\mathbf{L} \in \mathcal{L}_n$ if it satisfies:

(P1) $0 \in \mathbf{L}$.

(P2) $\mathbf{L} \subseteq [0,1]$.

(P3) For any sequence $L_1, \dots, L_{n-1} \in \mathbf{L}$, let $L^* = \alpha \left(\sum_{i=1}^{n-1} L_i + N \right)$. If $L^* \leq 1$, then $L^* \in \mathbf{L}$.

Given N, α, n, there may be more than one set $\mathbf{L} \in \mathcal{L}_n$ (take the interval $[0,1]$ and the set of rationals in this interval as two trivial members of this class). However, we will show that there exists an element of class $\mathcal{L}_n$ which is finite, and that this set is the required support for the power levels such that the theorem holds.

Lemma 1. If $\mathbf{L}^1 \in \mathcal{L}_n$ and $\mathbf{L}^2 \in \mathcal{L}_n$, then $\mathbf{L}^1 \cap \mathbf{L}^2 \in \mathcal{L}_n$.

Proof. For $\mathbf{L}^1 \cap \mathbf{L}^2$, properties (P1) and (P2) are trivial. To check (P3), choose a sequence $L_1, \dots, L_{n-1} \in \mathbf{L}^1 \cap \mathbf{L}^2$ and define L^* as in (P3). Then, clearly $L^* \in \mathbf{L}^1$ and $L^* \in \mathbf{L}^2$ by the assumption that $\mathbf{L}^1 \in \mathcal{L}_n$ and $\mathbf{L}^2 \in \mathcal{L}_n$, so $L^* \in \mathbf{L}^1 \cap \mathbf{L}^2$. □

Note that this proof generalizes easily to arbitrary intersections. Then, we can define a minimal set $\mathbf{L}_n^*$ such that $\forall \mathbf{L} \in \mathcal{L}_n$, $\mathbf{L}_n^* \subseteq \mathbf{L}$. The proof of this only requires the (non-constructive) definition $\mathbf{L}^* = \bigcap \{\mathbf{L} : \mathbf{L} \in \mathcal{L}_n\}$.

Now, given any $\mathbf{L} \in \mathcal{L}_n$, define the floor quantization $q(x)$ as the maximum $L_k \in \mathbf{L}$ s.t. $L_k \leq x$. Equivalently, $q(x) = \sum_k L_k I_{\{x \in [L_k, L_{k+1})\}}$

For a vector of numbers $\mathbf{x} = \{x_1, \dots, x_n\} \in [0,1]^n$, define $q(\mathbf{x})$ as the term by term quantization $q(\mathbf{x})_i = q(x_i)$. Note that the support for the quantized power level distribution is now $\mathbf{L}$.

Now, assume $\mathbf{x} \in \mathcal{C}^n$. We will show that $q(\mathbf{x}) \in \mathcal{C}^n$. Hence, rounding down to $\mathbf{L}$ does not destroy capture events and thus that capture probability cannot decrease when considering the set of levels $\mathbf{L}$ if $\mathbf{L} \in \mathcal{L}_n$.

Now, since $\mathbf{x} \in \mathcal{C}^n$, this means that $x_n^{(n)} \geq \alpha \left(\sum_{j=1}^{n-1} x_j^{(n)} + N \right)$. Let $L_j = q(x_j^{(n)})$. Clearly, then, $x_n^{(n)} \geq \alpha \left(\sum_{j=1}^{n-1} L_j + N \right)$

Since $\mathbf{L} \in \mathcal{L}_n$, $\exists L_i \in \mathbf{L} : L_i = \alpha \left(\sum_{j=1}^{n-1} L_j + N \right)$. So, $x_n^{(n)} \geq L_i$ and hence $q(x_n^{(n)}) \geq L_i$. Then, $q(\mathbf{x}) \in \mathcal{C}^n$ since $q(x_n^{(n)}) \geq \alpha \left(\sum_{j=1}^{n-1} q(x_j^{(n)}) + N \right)$.

Lemma 2. The minimal set $\mathbf{L}_n^*$ is finite.

Proof. Consider the set of polynomials of the form:

$$\mathbf{P} = \{P(\alpha) : P(\alpha) = N\alpha(b_0 + b_1\alpha + b_2\alpha^2 + \ldots + b_K\alpha^K),\ b_j \in \mathbb{N},\ P(\alpha) \leq 1\} \qquad (8)$$

Clearly $\mathbf{P}$ is finite: the maximum order of any polynomial in $\mathbf{P}$ cannot be s.t. $N\alpha^K > 1$, and no coefficient can be such that $Nb_j\alpha^j > 1$, $1 \leq j \leq K$. Also, $\mathbf{P} \in \mathcal{L}_n$: If $P_1(\alpha), \ldots, P_{n-1}(\alpha) \in \mathbf{P}$, let $P^*(\alpha) = \alpha\left(\sum_{i=1}^{n-1} P_i(\alpha) + N\right) = \alpha\left(\sum_{i=1}^{n-1} \alpha N b_i(\alpha) + N\right) = \alpha N b'(\alpha)$ where $b'(\alpha)$ is also a polynomial in α. So, if $P^*(\alpha) \leq 1$, $P^*(\alpha) \in \mathbf{P}$ by definition.

Thus, since $\mathbf{L}_n^* \subseteq \mathbf{P}$ by definition, and $\mathbf{P}$ is finite, then $\mathbf{L}_n^*$ is finite. □

Thus, we have a finite set $\mathbf{L}_n^*$ for which capture probability $\Pr(\mathcal{C}^n; F_t)$ is not decreased when the support for F_t is rounded to a discrete support $\mathbf{L}_n^*$ resulting in a probability vector ϕ_t, thus proving the theorem. Furthermore, $\mathbf{P}$ provides a support for ϕ_t which is independent of n. Thus, for the cases when n is not known, it is enough to restrict ourselves to find an optimal probability vector ϕ_t defined on $\mathbf{P}$. □

5 Stability limits

We consider now the achievable performance of the proposed dynamic control schemes. The results in this section closely parallel the ones in [2], where a general model of Slotted Aloha with multipacket reception and capture is considered but where capture probabilities C_n were fixed and thus invariant with respect to control actions. We will focus on studying the achievable throughput when the number of backlogged users n_t is known. Although this is not a realistic situation, it serves as an upper bound for the achievable throughputs when n_t is unknown. Also, we will concentrate on the infinite user case $M = \infty$.

Recall that the drift of a Markov process $\{n_t, t \in \mathbb{N}\}$ at state k (where the state space is over the natural numbers) is defined as $D_k = E(n_{t+1} - n_t | n_t = k)$. For an infinite population Aloha model, drift analysis allows us to study the process stability (ergodicity) and to get upper bounds on the achievable throughput. Heuristically speaking, the maximum achievable throughput will be the maximum input arrival rate for which n_t is a stable, ergodic chain. Two results will be useful:

- *Pakes' Lemma [6]*: If for an irreducible, aperiodic Markov chain n_t, $\exists K$ s.t. $\forall k > K, D_k < -\epsilon < 0$ for some constant (independent of k) $\epsilon > 0$, then the Markov chain is ergodic.
- Conversely (Kaplan [3]), if $n_{t+1} - n_t$ is bounded below and $\exists K'$ s.t. $D_k \geq 0\ \forall k > K'$, then the Markov chain n_t is unstable.

So, the stability of the backlog process is determined by the drift behavior for large values of n_t in case of an infinite user population. Since $n_{t+1} - n_t = A_t - I_{\mathcal{C}_t}$, we have that $D_k = \lambda - P_k$, where $P_k = \Pr(\mathcal{C}_t | n_t = k) = \sum_{n=0}^{\infty} C_n(\phi_t) \Pr(f_t = n | n_t = k)$.

As $k \to \infty$, the backlog traffic R_t conditioned on $n_t = k$ will converge in distribution to a Poisson random variable with parameter kq_t, and thus, for large backlog values, the offered traffic to the network will be approximately Poisson with parameter $\lambda + kq_t$:

$$P_k(\phi_t, q_t) = \sum_{n=0}^{\infty} C_n(\phi_t) e^{-\lambda - q_t k} \frac{(\lambda + q_t k)^n}{n!}$$

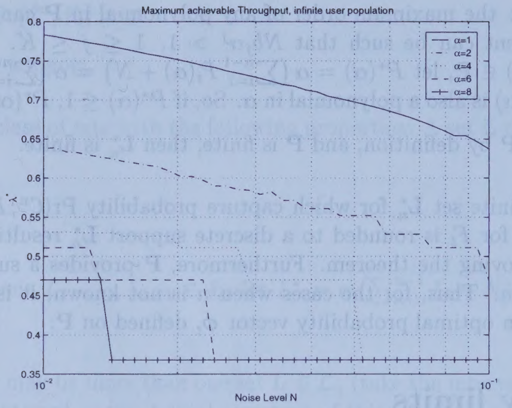


Figure 1: Maximum achievable λ^* from Eq. 9.

If we know $n_t = k$, then the vector $\boldsymbol{\pi}_t$ degenerates to $\pi_t(i) = \delta_{ik}$, so Eq. 2 becomes $(q_t^*, \boldsymbol{\phi}_t^*) = \operatorname{argmax} P_k(\boldsymbol{\phi}_t, q_t) = \operatorname{argmin} D_k(\boldsymbol{\phi}_t, q_t)$, where then $D_k(\boldsymbol{\phi}_t, q_t) = \lambda - \sum_{n=0}^{\infty} C_n(\boldsymbol{\phi}_t) e^{-\lambda - q_t k} \frac{(\lambda + q_t k)^n}{n!}$
Let $D_\lambda(\boldsymbol{\phi}, x) = \lambda - \sum_{n=0}^{\infty} C_n(\boldsymbol{\phi}) e^{-\lambda - x} \frac{(\lambda + x)^n}{n!}$, so then $D_k = D_\lambda(\boldsymbol{\phi}, q_t k)$.

$$\lambda^* = \sup \left\{ \lambda \in [0,1] : \inf_{\boldsymbol{\phi} \in \Delta_m, x \geq 0} D_\lambda(\boldsymbol{\phi}, x) < 0 \right\} \tag{9}$$

We can then readily establish the following result, which is a direct extension of [2, Thms. 2 and 3]:

Lemma 3. For input arrival rates $\lambda < \lambda^*$ there exist retransmission and power control strategies for which the system is stable if n_t is known. For arrival rates $\lambda > \lambda^*$, the system is unstable for any control algorithm.

Proof. If $\lambda < \lambda^*$, then there exists $(\boldsymbol{\phi}', x')$ such that $D_\lambda(\boldsymbol{\phi}', x') < 0$. Hence, if $n_t = k$, by using $\boldsymbol{\phi}'$ and $q_t = x'/k$ as controls then $D_k = D_\lambda(\boldsymbol{\phi}', x') < 0$. Thus, the system would always be stable by Pakes' lemma. By using the optimal actions $(\boldsymbol{\phi}^*, q_t^*)$ from Eq. 2 which minimize drift, $D_\lambda(\boldsymbol{\phi}^*, q_t^* k) \leq D_\lambda(\boldsymbol{\phi}', x') < 0$ for all k large, so stability is preserved.

If $\lambda > \lambda^*$, then for any control pair $(\boldsymbol{\phi}, x)$ we will have that $D_\lambda(\boldsymbol{\phi}, x) \geq 0$, so the system will be unstable by Kaplan's lemma (as clearly $n_{t+1} - n_t \geq -1$). ☐

Even though Eq. 9 provides no closed form solution for λ^* since it is an implicit equation in λ, it is still possible to numerically evaluate Eq. 9 for specific values of α and N which will provide upper bounds on the real system performance where n_t is not known. Figure 1 plots Eq. 9 for several values of N and α.

6 Numerical Results

Numerical simulations were carried out in order to explore the performance (throughput and mean backlog) of the proposed system. Two system setups were considered: A finite population model with $M = 50$ and an infinite population one. For the latter case, truncation of vector $\boldsymbol{\pi}_t$ is necessary, so only entries above 10^{-6} were kept in order to preserve numerical accuracy without incurring excessive storage requirements. The noise levels studied (N) varied between 10^{-2} and 10^{-1} (SNR of 10 to 20 dB) with several threshold constants α.

In order to convey an idea of the size of the numerical optimization involved in performing the control indicated by Eq. 2, Table 1 lists the number of levels $m = \|\mathcal{L}\|$ used for optimization according to the construction of Eq. 8. Since the online algorithm optimizes both power level probabilities and retransmission probability, the total number of parameters for the numerical optimization algorithms is $m + 1$.

	α		
N	2	4	8
0.01	51	10	4
0.1	6	3	2

Table 1: Optimal number of power levels used for system control, from Eq. 8.

Figure 2 presents the measured throughput and mean backlog for two different system configurations. These results were obtained by averaging over 10^5 slots. The figure on the left plots the performance as a function of $\lambda = Mq_a$ for particular values of $N = 0.01$ and $\alpha = 2$ with a finite user population of $M = 50$, for both feedback models presented before as well as for the case when n_t is known. The figure on the right plots, for $M = \infty$, the mean backlog as a function of N and α for the fixed value of $\lambda = 0.36$, which is known to be a stable value for any combination of parameters such that $\alpha N \leq 1$. For this case, the throughput obtained in all cases was approximately equal to λ.

It can be seen that there is very little difference in terms of throughput for both feedback models but there is an increase in the backlog (and hence in expected delay) when only ternary feedback is available. There is also a practically imperceptible difference between the cases when n_t is known and when only f_t is available. However, this is an idealized situation, as we would normally have imperfect estimates of f_t in practice.

Also, it can be seen that the upper bound for the throughput presented in Section 5 is tight, as it is achieved even when n_t is not known. There is also, as expected, a deterioration in the mean backlog when α is increased.

7 Conclusions and Future Work

The proposed dynamic control scheme can provide important gains in throughput by exploiting the capture phenomenon at the receiver. These gains become more important if the system is designed as to have a small SIR threshold α. The main drawback of this approach is its high computational complexity and the assumption that the channel conditions and dynamics are perfectly known. Thus, simpler suboptimal and adaptive control policies may be required.

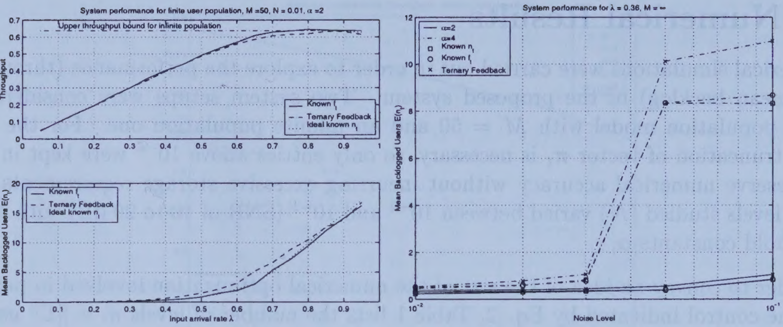


Figure 2: Performance as a function of system parameters λ, N and α.

Finally, it would be interesting to extend the current models to other, more complex, random access protocols.

References

[1] T. L. Fine, S. B. Wicker, T. Berger, and J. Halpern. Sensor-assisted aloha for wireless networks. In *Proceedings. 1998 IEEE International Symposium on Information Theory*, page 484, 1998.

[2] S. Ghez, S. Verdú, and S. Schwartz. Optimal decentralized control in the random access multipacket channel. *IEEE Transactions on Automatic Control*, 34(11):1153–1163, Nov. 1989.

[3] M. Kaplan. A sufficient condition for nonergodicity of a markov chain. *IEEE Transactions on Information Theory*, IT-25(4):470–471, July 1979.

[4] R. O. LaMaire, A. Krishna, and M. Zorzi. On the randomization of transmitter power levels to increase throughput in multiple access radio systems. *Wireless Networks*, 4(3):263–277, 1998.

[5] C. Namislo. Analysis of mobile radio slotted aloha networks. *IEEE Journal on Selected Areas in Communications*, SAC-2(4):583–588, July 1984.

[6] A. G. Pakes. Some conditions for ergodicity and recurrence of markov chains. *Operations Research*, 17:1058–1061, 1969.

[7] V. Wong and C. Leung. A transmit power control scheme for improving performance in a mobile packet radio system. *IEEE Transactions on Vehicular Technology*, 43(1):174–180, Jan. 1994.

[8] M. Zorzi and R. R. Rao. Capture and retransmission control in mobile radio. *IEEE Journal on Selected Areas in Communications*, 12(8):1289–1298, Oct. 1994.

Adaptive Coding with Pilot Signals.

Muriel Médard[1], Ibrahim Abou-Faycal[1], Upamanyu Madhow[2]
[1] Laboratory for Information and Decision Systems, Massachusetts Institute of Technology
[2] Electrical and Computer Engineering Department, University of California Santa Barbara

Abstract

We consider adaptive modulation and coding for channels with pilot symbol assisted modulation. Unlike most adaptive transmission schemes, we do not have any channel side information at the sender. The sender adapts not to the state of the physical channel traversed by its transmitted signal, but to the quality of the measurement afforded by pilot symbols at the receiver. We consider the case where no pilot symbols are used, where pilot symbols are used with adaptive reception but without adaptive transmission, and the case where both reception and transmission are done adaptively (while maintaining the average per symbol energy constant). Our results are in terms of capacity for binary signaling. We show that, under appropriate optimizing conditions for the spacing between consecutive pilot symbols, pilot symbol assisted modulation is helpful in terms of capacity. Moreover, adaptation at both the sender and the receiver yields clear advantages over adaptation at the receiver only.

1 Introduction

In order to make best use of wireless resources, many different types of adaptive schemes are employed. Adaptive schemes seek to modify the transmission scheme used by the sender according to the state of the channel seen by the receiver. Generally, such schemes involve feedback, concerning the state of the channel, from the receiver to the sender. In an information-theoretic context, adaptive signaling is used for Markov channels with perfect sender and receiver channel side information [Wol78, GV97] or imperfect channel side information [CS99, MS00, Vis99, Kle01]. Power control is a commonly used type of adaptive transmission. Many practical schemes consider modifying the modulation used in order to combat fading. A common means of adapting transmission is to use different types of modulation [TH99, QC99], for instance different levels of QAM constellations [WH00, WS95, SKM94, TH97, GC97, LJ98, WYH00, THK99], according to the state of the channel and possibly other considerations such as multi-user interference [GA99].

In this paper, we consider a different type of adaptive scheme. While the scheme still adapts the transmission to the channel seen by the receiver, the transmitter does not use feedback to determine its policy. Instead, the transmitter takes into account the time-varying quality of the channel measurement available at the receiver in order to modify its transmission policy. Thus, the transmitter adapts to the quality of the channel measurement, rather than to the quality of the channel (in terms of carrier to noise ratio or other metric.)

Our model is the following. We consider a single sender and receiver, connected by a time-varying Rayleigh fading channel. The Rayleigh fading channel is modeled as a Gauss-Markov process. The sender transmits coded and modulated data. The sender has no information regarding the state of the channel and, therefore, does not adapt its transmission scheme in response to fades. Moreover, at regular intervals, the sender transmits a constant and known pilot symbol, whose purpose is to enable measurement of the channel at the receiver. The sender thus transmits coded data, periodically interrupted by pilot symbols. Such a scheme is often referred to as pilot symbol assisted modulation [TAG99, KKJ$^+$97, Cav91, SS93, TH95]. Pilot symbols are commonly used to improve detection [HK96, MB84] and decoding [WC99]. The pilot symbols have energy equal to the average energy constraint. The only channel estimate available at the receiver comes from the pilot symbols, hence there is no data-directed estimation of the channel. The channel estimate at a time sample k depends on the received pilot symbols through its position with respect to the pilot symbols. Given the estimate of the channel obtained through the pilot symbols, the channel is no longer Rayleigh but Ricean with a known specular component. We consider binary signaling, since such signaling performs well at low SNRs and achieves capacity for low SNRs for Rayleigh channels. We also consider a discrete time, sampled system.

Our purpose is to asses the benefits of performing adaptive transmission which takes into account the quality of the estimate obtained at the receiver. We seek to maximize mutual information. The maximization of mutual information subsumes the optimization of both the modulation and the coding. We consider three cases. First, we consider the case where no pilot symbols are transmitted and the channel is not estimated at the receiver. The channel at the receiver is then a Rayleigh channel. Since we do not perform any data-directed estimation, the capacity of the system is the same as if all the channel samples were mutually independent (by averaging over very long times). The capacity of such channels was found in [AFTS97]. For low to moderate SNRs, binary signaling is optimal. The purpose of considering the Rayleigh channel with no pilot symbols is to establish a basis of comparison for the other two cases, which do employ pilot symbols. Indeed, the first question we pose is whether it is preferable to forego pilot symbols and channel estimation altogether and devote to coded data the time allocated to pilot symbols.

Second, we consider the case where we use pilot symbols to aid in the detection and decoding at the receiver but do not modify the distribution of the transmitted signal. We term this scheme the non-adaptive scheme with pilot tones, since the scheme is adaptive at the receiver but not at the sender. The sender uses the distribution that is optimal for transmission when the channel is block-faded. In effect, the sender behaves as though the channel estimate did not vary between pilot symbols. Since we look at mutual information, we may consider this case to be an upper bound to the case where neither the distribution nor the coding is modified according to the distance between the transmitted signal and the pilot symbols. Alternatively, we may view our results as representing capacity when the distribution of the transmitted signal is fixed.

Third, we consider the case where we use pilot symbols and adaptive transmission, but maintain the average per symbol power constant in the coded data. We consider the case where the estimation of the channel is causally performed using the last pilot symbol, and the case where the estimation is non-causally based on the last transmitted pilot along with the pilot transmitted next. At each time sample, the sender changes his transmission policy according to the distance to pilot symbols. We compare the performance of the second and third cases in order to establish the benefit of adaptive transmission.

In Section 2, we present our channel model and the principles of non-causal and causal estimation. In Section 3, we discuss non-adaptive and adaptive schemes used. In Section 4, we present our numerical results. These results allow us to optimize numerically the spacing between pilot symbols for adaptive and non-adaptive schemes, since such an optimization cannot be obtained in closed form. Using the optimized spacing between pilot symbols for the different schemes, we can evaluate the benefit of adaptive schemes. Finally, conclusions and directions for future work are presented in Section 5.

2 Channel Model

We consider the following discrete-time model for the Rayleigh fading channel

$$Y = AX + N,$$

where X is the channel input, Y the output, and A and N are independent complex circular Gaussian random variables with mean zero and variance σ_A^2 and σ_N^2 respectively. Equivalently, the amplitude of the fading coefficient A is Rayleigh distributed and its phase is uniform. The input X is average power limited: $E\left[|X|^2\right] \leq P$.

We assume that the fading process is a first-order Gauss-Markov process

$$A_k = \alpha A_{k-1} + Z_{k-1},$$

where k denotes the time index. Depending on the scale of the time variations of the channel, α[1] may vary between one (corresponding to no channel variations) and zero (where the fading is fast, changing independently from one symbol time to the next.) A similar channel model was introduced in [MÓ0], where the relation between this channel and the coherence time is established.

For a coherence time of T_c and transmission over a bandwidth of W, α is determined by $\alpha^{T_c W} = \phi$ where ϕ is the level of decorrelation we deem necessary in our definition of coherence time. In the literature, the correlation coefficient is taken to vary from 0.9 ([CL75]) to 0.37 ([BN63]) for a time separation of T_c. For bandwidths in the 10 kHz range, and Doppler spreads of the order of 100 Hz, α will typically range between 0.9 and 0.99. For instance, for a Doppler spread of 100 Hz, $W = 10^4$ Hz, $\phi = 0.1$, we have $\alpha = 0.977$. For a Doppler spread of 200 Hz, $\alpha = 0.955$. For a Doppler spread of 50 Hz, $\alpha = 0.989$.

When no pilot signals are used and the correlation between fading coefficients is ignored, the channel behaves as a memoryless Rayleigh fading channel. The capacity of this channel was studied in [AFTS97] and the optimizing input distribution was found to be discrete with a finite number of mass points one of then located at the origin. Furthermore, a binary distribution was found to be optimal at low and moderate values of the SNR motivating the use of binary input strategies in this paper.

For an adaptively coded system, we consider the case where a pilot signal of power P is transmitted once at the beginning of each interval of length T, enabling the receiver to estimate the fading coefficients governing the channel statistics. The resulting channel outputs are

$$Y_{lT} = A_{lT}\sqrt{P} + N_{lT}, \quad l \in \mathbb{N}.$$

Based on the observation of $\{y_{lT}\}_l$, estimates of the fading coefficients $\{A_k\}_{lT+1}^{(l+1)T-1}$ are obtained and at each time step k, using its estimate the receiver experiences a Ricean channel where the fading coefficient has mean $\hat{A}_k$ and variance v_k, where $\hat{A}_k$ is the estimate of the value of A_k and v_k is the variance of the estimation error:

$$p_{Y_k|X_k}(y_k|x_k) = \frac{1}{\pi(v_k|x_k|^2 + \sigma_N^2)} \exp\left\{ -\frac{|y_k - \hat{A}_k x_k|^2}{v_k|x_k|^2 + \sigma_N^2} \right\}, \quad lT < k < (l+1)T.$$

---

[1] Since the fading is Rayleigh distributed, it is sufficient to consider only real positive values for the parameter α.

Therefore[2], for a given received sequence $\{y_{lT}\}_l$ the mutual information $I_k(X_k;Y_k|\{Y_{lT}\} = \{y_{lT}\})$ at time $k \in \{lT+1,\cdots,(l+1)T-1\}$ is given by

$$\begin{aligned} I_k(X_k;Y_k|\{Y_{lT}\} = \{y_{lT}\}) =& p_k(1)\int p_{Y_k|X_k}(y|x_k(1))\ln\frac{p_{Y_k|X_k}(y|x_k(1))}{p_{Y_k}(y)}dy \\ &+ (1-p_k(1))\int p_{Y_k|X_k}(y|x_k(2))\ln\frac{p_{Y_k|X_k}(y|x_k(2))}{p_{Y_k}(y)}dy, \end{aligned}$$

where the probability mass function of the binary input $\{x_k(1),x_k(2)\}$ is $\{p_k(1),(1-p_k(1))\}$.

Furthermore, since we are eliminating the possibility of data-directed estimation, the achievable rates[3] are hence given by

$$\mathsf{E}\left[\frac{1}{T}\sum_{k=lT+1}^{(l+1)T-1} I_k(X_k;Y_k|\{Y_{lT}\} = \{y_{lT}\})\right] = \frac{1}{T}\sum_{k=lT+1}^{(l+1)T-1}\mathsf{E}\left[I_k(X_k;Y_k|\{Y_{lT}\} = \{y_{lT}\})\right], \tag{1}$$

where the expectation is over the (correlated) random variables $\{Y_{lT}\}$ which are circular Gaussian distributed with mean zero and variance $\sigma_A^2 P + \sigma_N^2$.

Overall we consider two estimation procedures:

1. The receiver performs a causal estimation of the channel parameters based on the most recently sent pilot tone. For each interval, based on the observation of y_{lT}, estimates of the fading coefficients $\{A_k\}_{lT+1}^{(l+1)T-1}$ are obtained by a standard Bayesian least square estimation procedure yielding

$$\hat{A}_k = \frac{\sqrt{P}\sigma_A^2}{P\sigma_A^2+\sigma_N^2}\alpha^{(k-lT)}y_{lT} \tag{2}$$

$$v_k = \sigma_A^2 - \frac{P\sigma_A^4}{P\sigma_A^2+\sigma_N^2}|\alpha|^{2(k-lT)}, \qquad lT < k < (l+1)T. \tag{3}$$

2. The estimation is non-causal and based on the most recently sent pilot tone along with the one transmitted next, i.e. $\hat{A}_k$ is based on the values of y_{lT} and $y_{(l+1)T}$ whenever $lT < k < (l+1)T$. In this case

$$\begin{aligned} \hat{A}_k =& \frac{\sqrt{P}\sigma_A^2}{(P\sigma_A^2+\sigma_N^2)^2 - P^2\alpha^{2T}}\Big[\left[\alpha^{(k-lT)}(P\sigma_A^2+\sigma_N^2) - P\alpha^{2T-(k-lT)}\right]y_{lT} \\ &+ \left[\alpha^{T-(k-lT)}(P\sigma_A^2+\sigma_N^2) - P\alpha^{T+(k-lT)}\right]y_{(l+1)T}\Big] \end{aligned} \tag{4}$$

$$v_k = \sigma_A^2 - \frac{P\sigma_A^4}{(P\sigma_A^2+\sigma_N^2)^2 - P^2\alpha^{2T}}\left[(P\sigma_A^2+\sigma_N^2)\left[\alpha^{2(k-lT)} + \alpha^{2[T-(k-lT)]}\right] - 2P\alpha^{2T}\right] \tag{5}$$

3 Non-Adaptive and Adaptive Coding

First we consider the scheme where the transmitter does not adapt its transmission strategy to the statistics of the channel estimates used at the receiver. More precisely, we compute the achievable rates when the transmitter is using a single fixed input distribution at all times. This of course will allows us to quantify the performance of a system when the transmitter does not adapt its coding strategy and uses instead one fixed codebook independently of the time index.

---

[2] Under the assumption of binary signaling.

[3] Note that the mutual information at times $k = lT$ is zero.

We look mainly at the case where the transmitter considers the channel to be block-faded where, the fading coefficient is assumed constant over intervals of length T and changing independently from one interval to the next. Consequently, we find first the optimal input distribution for the block-faded system, and then compute the average mutual information under this distribution for different values of α. While when $\alpha = 1$ the model based on the first estimation procedure would correspond to block-fading, as α decreases we expect the performance of this non-adaptive system to deteriorate, and we quantify this performance loss.

Here we consider the scheme where, without any channel state information, the transmitter takes into consideration the statistics of the channel estimates at the receiver. It adapts accordingly its modulation and coding to maximize the rates that can be reliably transmitted over the channel. While no optimal power allocation is performed here (a constant amount of power is used instead,) at each time step the transmitter uses a "good" codebook achieving the highest mutual information of the Ricean channel the receiver sees.

Equivalently, one can think of the problem as that of finding the best input strategy that maximizes the expected mutual information $\mathsf{E}\left[I_k\big(X_k; Y_k|\{Y_{lT}\} = \{y_{lT}\}\big)\right]$ for each time step between $lT + 1$ and $(l + 1)T - 1$. For these computations, we considered the two estimation methods described in section 2, i.e. for the pair $\Big(\{\hat{A}_k\}, \{v_k\}\Big)$ given by either equations (2) and (3), or equations (4) and (5).

We use standard matlab tools to optimize, for each time period $k \in \{1, \cdots, T-1\}$, the expected mutual information over the input probability distribution. The corresponding optimal distribution yields of course the highest achievable rates depending on how far the transmission is occurring with respect to the pilot signals. Sending pilot tones frequently clearly reduced the rates as a significant portion of the time and power is used to estimate the channel and no information is conveyed from the transmitter to the receiver. On the other hand, when the pilots are used very infrequently, the channel estimates at the receiver are of poor quality and the information rates are low. One of the questions that we will try to answer in the following section is: what is the optimal value of T that would yield the best compromise?

4 Numerical Results

In figures 1 and 2 we have drawn the achievable rates for different values of the SNR, for $\alpha = 0.95$ and 0.99 respectively. In these plots the value of T varies between 2 and 30. The dashed lines in these plots represent the non-adaptive scheme, while the solid lines are those of the adaptively coded system, when the causal estimation method is used.

As is apparent from figure 1, in the considered ranges of the SNR, there is no benefit gained from using pilot symbol assisted modulation for $\alpha = 0.95$. Indeed, for these low correlation models, not sending a pilot tone and coding instead to a memoryless Rayleigh fading channel outperforms the adaptive and non-adaptive causal scheme we have described. However, for $\alpha = 0.99$, figure 2 shows not only that an improvement is possible, but also that there is a trade-off between infrequent transmission of pilots and the channel estimate quality. Indeed, for every value of the SNR there is an optimal value for T where the rates are maximized, and these optimal values are apparent in the figure. This holds for both the adaptive and non-adaptive schemes and the relative improvement that can be achieved with the adaptive scheme considered ranges between 4.2% and 7.5% depending on the SNR. Moreover, as expected, the

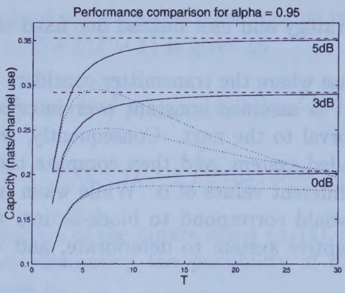


Figure 1: Best achievable rates function of T and the SNR for $\alpha = 0.99$

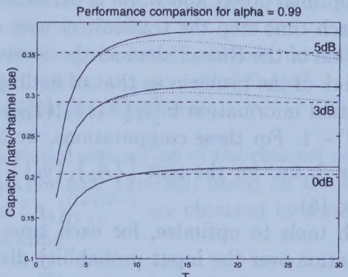


Figure 2: Best achievable rates function of T and the SNR for $\alpha = 0.99$ for adaptive and non-adaptive coding for causal estimation

adaptive scheme outperforms the non-adaptive one by 1.5% to 3.2% when the SNR ranges between 0 and 5dB. This gain becomes more significant for smaller values of α as it can be seen in figure 1. For illustrative purposes, in figures 3, 4 and 5 we have drawn the achievable rates for given values of the SNR, as a function of α when T varies between 2 and 30.

In figures 6 and 7, we have drawn the achievable rates for the same values of the SNR, α and T, when the non-causal estimation method is used. Figure 8 shows the achievable rates for 3dB SNR, for different values of α. As expected, the non-causal scheme outperforms the causal one. We note that the optimal values of T appear to be smaller. Initially, one would expect the opposite since the quality of the channel estimates for large values of T is significantly better especially in the mid-interval range. This suggests that choosing a large value of T is beneficial, especially that it reduces the fraction of time when the channel is idle. However, this is not the case. Intuitively speaking, the gain one gets from the two-pilots based channel estimation is considerable making the more frequent transmission of pilots here significantly better.

5 Conclusions and Future Work

We have investigated the use of adaptive and non-adaptive sender strategies for time-varying channels with pilot symbol assisted modulation and no feedback from the receiver to the

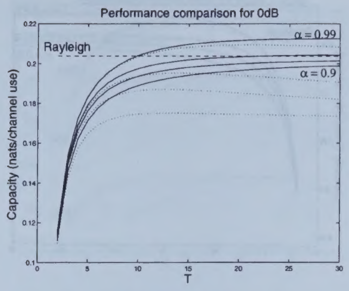


Figure 3: Best achievable rates at 0dB function of T when $\alpha = 0.9, 0.95, 0.97$ and 0.99 for adaptive and non-adaptive coding for causal estimation

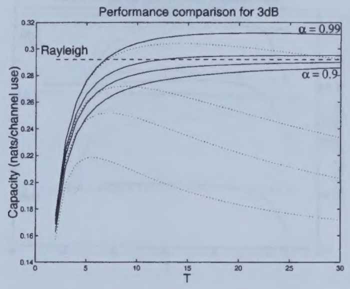


Figure 4: Best achievable rates at 3dB function of T when $\alpha = 0.9, 0.95, 0.97$ and 0.99 for adaptive and non-adaptive coding for causal estimation

sender. We have shown that, depending on the rate of change of the channel, adaptive sender strategies, properly optimized for spacing between consecutive pilot tones, can improve channel capacity.

Our adaptive strategy maintains average power constant on a symbol-per-symbol basis. A natural extension is to investigate to what extent adaptive power modulation can further improve capacity. Establishing an optimal power allocation is difficult, however, because we do not have a closed form for capacity. Preliminary results on some heuristics methods, akin to waterfilling, which attempt to maintain the interference constant, do not generally lead to improvements over constant power allocations.

References

[AFTS97] I.C. Abou-Faycal, M.D. Trott, and S. Shamai (Shitz). The capacity of discrete-time Rayleigh fading channels. *ISIT*, 8(2):473, 1997.

[BN63] P.A. Bello and B.D. Nelin. The effect of frequency selective fading on the binary error probabilities of incoherent and differentially coherent matched filter receivers. *IEEE Transactions*, CS-11(2), June 1963.

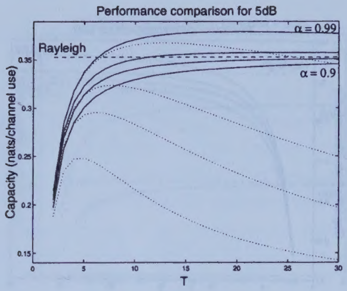


Figure 5: Best achievable rates at 5dB function of T when $\alpha = 0.9, 0.95, 0.97$ and 0.99 for adaptive and non-adaptive coding for causal estimation

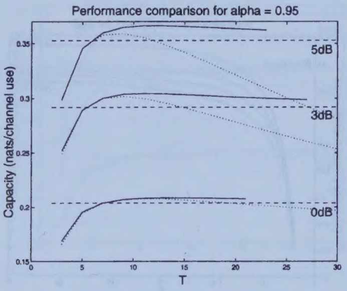


Figure 6: Best achievable rates function of T and the SNR for $\alpha = 0.95$ for adaptive and non-adaptive coding for non-causal estimation

[Cav91] J.K. Cavers. An analysis of pilot symbol assisted modulation for Rayleigh fading channels. *IEEE Transactions on Vehicular Technology*, 40:686–693, November 1991.

[CL75] D.C. Cox and R.P. Leck. Correlation bandwidth and delay spread multipath propagation statistics for 910 mhz urban mobile radio. *IEEE Transactions*, COM-23(1), November 1975.

[CS99] G. Caire and S. Shamai. On the capacity of some channels with channel state information. *IEEE Transactions on Information Theory*, 45(6):2007–2019, September 1999.

[GA99] A.J. Goldsmith and M.-S. Alouini. Area spectral efficiency of cellular mobile radio systems. *IEEE Transactions on Vehicular Technology*, 48(4):1047–1066, July 1999.

[GC97] A. Goldsmith and S.G. Chua. Variable rate variable power MQAM for fading channels. *IEEE Transactions on Communications*, 45:1218–1230, October 1997.

[GV97] A.J. Goldsmith and P. Varaiya. Capacity of fading channels with channel side information. *IEEE Transactions on Information Theory*, November 1997.

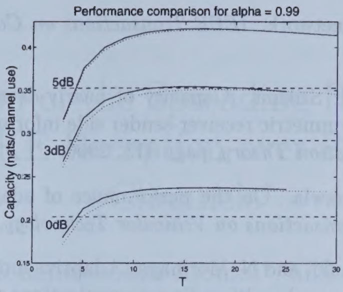


Figure 7: Best achievable rates function of T and the SNR for $\alpha = 0.99$ for adaptive and non-adaptive coding for non-causal estimation

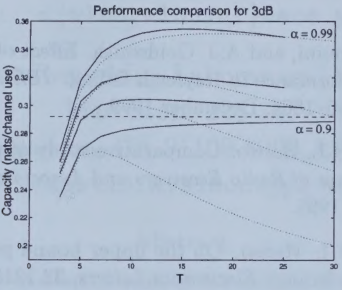


Figure 8: Best achievable rates at 3dB function of T when $\alpha = 0.9, 0.95, 0.97$ and 0.99

[HK96] P. Ho and J.H. Kim. Pilot symbol-assisted detection of cpm schemes operating in fast fading channels. *IEEE Transactions on Communications*, 44:337–347, March 1996.

[KKJ+97] Y.S. Kim, C.J. Kim, G.Y. Jeong, Y.J. Bang, H.K. Park, and S.S. Choi. New Rayleigh fading channel estimator based on PSAM channel sounding technique. *icc*, pages 1518–1520, June 1997.

[Kle01] T.E. Klein. *Capacity of Gaussian Noise Channels with Side Information and Feedback.* MIT Ph.D. Thesis, February 2001.

[LJ98] C.H. Lim and J.K. Jeong. Adaptive modulation using the multipath fading compensation. *IEEE Transactions on Vehicular Technology*, 34:940–942, 1998.

[MÓ0] M. Médard. The effect upon channel capacity in wireless communicationsof perfect and imperfect knowledge of the channel. *IEEE Transactions on Information Theory*, 46(3):935–946, 2000.

[MB84] J.P. McGeehan and A.J. Bateman. Phase locked transparent tone-in-band (TTIB): a new spectrum configuration particularly suited to the transmission of data over

SSB mobile radio network. *IEEE Transactions on Communications*, COM-32:81–87, January 1984.

[MS00] M. Médard and R. Srikant. Capacity of nearly-decomposable markovian fading channels under asymmetric receiver-sender side information. In *International Symposium on Information Theory*, page 413, 2000.

[QC99] X. Qiu and K. Chawla. On the performance of adaptive modulation in cellular systems. *IEEE Transactions on Vehicular Technology*, 48(4):1237–1251, July 1999.

[SKM94] S. Sampei, S. Komaki, and N. Morinaga. Adaptive modulation/ TDMA scheme for large capapcity personal multimedia communications systems. *IEICE Transactions on COmmunications*, E77-B:1096–1103, September 1994.

[SS93] S. Sampei and T. Sunaga. Rayleigh fading compensation for qam in land mobile radio communications. *IEEE Transactions on Vehicular Technology*, 42:137–147, May 1993.

[TAG99] X. Tang, M.-S. Alouini, and A.J. Goldsmith. Effect of channel estimation error on M-QAM BER performance in Rayleigh fading. *IEEE Transactions on Communications*, 47(12):1856–1864, December 1999.

[TH95] J. M. Torrance and L. Hanzo. Comparative study of pilot symbol assisted modem schemes. *Proceedings of Radio Receivers and Associated Systems Conference*, pages 36–41, September 1995.

[TH97] J.M. Torrance and L. Hanzo. On the upper bound peformance of adaptive QAM in a slow Rayleigh fading. *Electronics Letters*, 32:1218–1230, October 1997.

[TH99] J. M. Torrance and L. Hanzo. Adaptive modems over slow indoor Rayleigh fading channels. *IEEE Transactions on Vehicular Technology*, 48(4):1237–1251, July 1999.

[THK99] J. M. Torrance, L. Hanzo, and T. Keller. Interference aspects of adaptive modems over slow Rayleigh fading channels. *IEEE Transactions on Vehicular Technology*, 48(5):1527–1535, September 1999.

[Vis99] H. Viswanathan. Capacity of time-varying channels with delayed feedback. *IEEE Transactions on Information Theory*, 45(2):761–771, March 1999.

[WC99] L. Welburn and J.K. Cavers. Pilot signals improve the performance of Reed-Solomon erros and erasures decoder in Rayleigh fading channels. *IEEE Transactions on Communications*, 47(5):1527–1535, May 1999.

[WH00] C.H. Wong and L. Hanzo. Upper-bound performance of a wide-band adaptive modem. *IEEE Transactions on Communications*, 48(3):367–369, March 2000.

[Wol78] J. Wolfowitz. *Coding Theorems of Information Theory.* Springer-Verlag, 1978.

[WS95] W.T. Webb and R. Steele. Variable rate QAM for mobile radio. *IEEE Transactions on Communications*, 43:2223–2230, July 1995.

[WYH00] C.H. Wong, B.L. Yeap, and L. Hanzo. Wideband burst-by-burst adaptive modulation with turbo equalization and iterative channel estimation. *Vehicular Technology Conference Proceedings*, 3:2044–2048, Spring 2000.

Optimization of the Spreading Bandwidth for Selective RAKE Reception over Rayleigh Fading Channels*

Wing C. Lau*, Mohamed-Slim Alouini*, and Marvin K. Simon◇

*Department of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, USA.
E-mail: <{wlau,alouini}@ece.umn.edu>

◇Jet Propulsion Laboratory (JPL)
Mail Stop: 238-343
Pasadena, CA 91109-8099, USA.
E-mail: <marvin.k.simon@jpl.nasa.gov>

Abstract

Building on the recent developments in the performance analysis of generalized selection combining (GSC), this paper examines the optimum spreading bandwidth for a fixed complexity GSC diversity receiver operating over independent identically distributed Rayleigh paths. For this purpose, the study considers three performance criteria, namely (i) average combined signal to noise ratio (SNR) at the GSC output, (ii) average bit error probability (BEP), and (iii) outage probability of the instantaneous combined SNR at the GSC output. For the average BEP criterion, results are presented for coherent combining. For the average combined SNR and some instances of the average BEP optimization problem, an accurate approximate estimate of this optimum bandwidth in the form of a solution of a transendantal equation is provided. In other cases, where the optimization is not easily tractable in an analytic fashion, a numeric-search procedure is used to find this optimum bandwidth for different performance criteria and system parameters of interest. Finally simplified rule of thumb type of formulas are also presented as a good reference for picking the optimum spreading bandwidth given a set of system parameters and a particular performance criterion of interest.

1 Introduction

Ultra wide-band transmission systems are expected to play a significant role in the deployment of next-generation wireless spread spectrum systems. One of the most promising features of such systems resides in the fact that increasing the spreading bandwith allows resolving additional multipaths resulting in an increased diversity which can be exploited by RAKE reception [1-3]. In this context, generalized selection combining (GSC) [4-20]

---

*This work was supported in part by the National Science Foundation Wireless Initiative Grant No. NSF/ECS-9979443 and in part by the National Science Foundation Grant CCR-9983462 .

recently received a great deal of attention as an alternative to conventional maximal-ratio combining (MRC), equal-gain combining, or selection combining (SC). The objective of GSC is to compromise between performance and implementation complexity. This is achieved by building a fixed complexity RAKE receiver (with a predetermined number of fingers L_c regardless of the number of available multipaths L) which continually combines, as per the rules of MRC for coherent systems or post-detection EGC for the noncoherent ones, the L_c paths with the highest instantaneous signal-to-noise ratio (SNR).

Previous work on this topic focused on the development of methodologies to derive exact closed-form expressions for various performance measures of GSC over Rayleigh and Nakagami fading channels. More specifically, the average output SNR of GSC was studied in [7, 9, 11, 12, 17] for independent identically distributed (i.i.d.) Rayleigh fading paths and in [8, 10] for non-i.i.d. Rayleigh fading paths. In addition, the GSC output SNR variance and outage probability over i.i.d. Rayleigh fading channels was derived in [9, 11, 12, 17] and in [17], respectively. The average symbol error probability for various modulations schemes was examined in [4, 5, 10, 13, 14, 15, 16, 17] for i.i.d. Rayleigh paths, in [4, 5, 6, 10, 17] for non-i.i.d Rayleigh paths, and in [18, 19] for i.i.d. Nakagami fading paths. Finally, an analysis of the effect of fading decorrelation on the average probability performance of GSC receivers was presented in [20]. Regardless of the performance measure or the type of fading, two main conclusions were obtained from the numerical results presented in these various studies. First, for a fixed number of available diversity paths L, most of the diversity gain is achieved by combining the strongest diversity paths and diminishing returns are obtained as the number of combined paths L_c increases. On the other hand, for a fixed number of combined paths L_c and for a fixed *average SNR per path* $\overline{\gamma}$ a significant performance improvement can be gained by increasing the number of available diversity paths L, which is, as pointed out in [12], proportional to the spreading bandwidth W_{ss} (more specifically $L \approx W_{\mathrm{ss}}\, \tau_{\max}$, where $\tau_{\max}$ is the channel maximum delay spread). Based on these conclusions and remarks, one may be tempted to conclude that, from a system design perspective, one should spread the transmitted signal as much as possible to increase the number L of available multipaths at the receiver reducing therefore the probability of deep fades on all paths and improving thereby the overall system performance. However, one has to recall that an increase in the number of available multipaths implies also a reduction in the average SNR per path for a fixed *total "captured" average SNR* $\overline{\gamma}_t$. Thus, in fact we are in the presence of a tradeoff and there is therefore an optimum choice, denoted by L^* for the number L of multipaths that should be made available to the RAKE receiver. In other words, for a fixed complexity RAKE receiver with L_c combined paths and operating over a frequency-selective channel with a known maximum delay spread $\tau_{\max}$ [s], there is an *optimum spreading bandwidth*, denoted by $W_{\mathrm{ss}}^* = L^*/\tau_{\max}$ [Hz], for which the GSC scheme performance is maximized.

The aim of this paper is to determine this optimum spreading bandwidth (or equivalently the optimum number of available paths L^*) for various performance criteria. Related studies examined the average bit error probability (BEP) optimization of conventional post-detection EGC receivers with noncoherent binary frequency-shift-keying (BFSK) detection [21, p. 549] and for conventional SC with noncoherent BFSK detection [22]; but, to the best knowledge of the authors this optimization has not yet been addressed for GSC receivers. For analytical tractability, several basic assumptions are made: (i) slow fading model, (ii) perfect channel estimation, and (iii) i.i.d. Rayleigh diversity paths. In addition, for completeness, we consider the optimum spreading band-

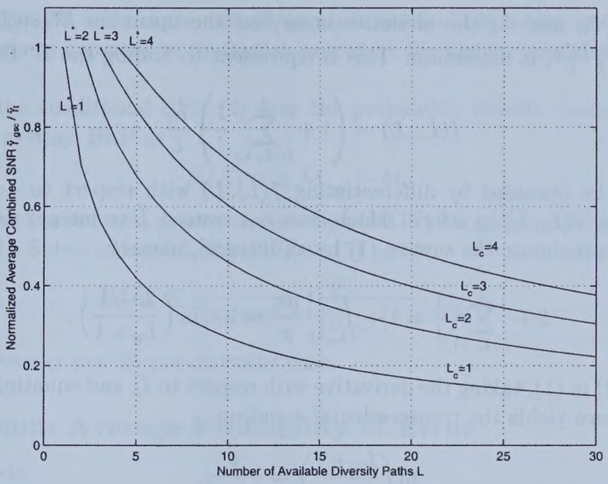


Figure 1: Normalized average combined SNR ($\overline{\gamma}_{\text{gsc}}/\overline{\gamma}_t$) as function of the number of available paths L for various values of the combined paths L_c.

width that maximizes three performance criteria: (i) average combined SNR at the GSC output, (ii) average probability of error, and (iii) outage probability of the instantaneous combined SNR at the GSC output (or equivalently instantaneous probability of error). For each of these performance criteria, we first fix the receiver complexity (i.e., L_c) and the total captured SNR then determine the optimum number of available paths that maximizes performance. We then look at the dual optimization problem where we still fix the receiver complexity but also the required performance (i.e. desired/satisfactory quality of service (QoS)) and rather seek to find the optimum number of available paths that minimizes the average total captured SNR, or equivalently the total average transmitted power. This is particularly important when the transmitter is a hand-held terminal with a limited lifetime battery and where the transmit power is a critical finite resource. This is also important for the downlink transmission from a base station to mobile units. Indeed in that case, although not that critical from an energy saving perspective, the optimization of the transmit power is rather motivated by a reduction in the interference level upon co-channel users.

The remainder of this paper is organized as follows. The next section determines the optimum spreading bandwidth that maximizes the output average SNR. Section 3 addresses the same question for a minimum average BEP criterion given a fixed total average SNR and then examines the dual problem when we rather seek to minimize the average total captured SNR for a fixed/required average BEP. Section 4 presents some results if the same optimization problem is posed for an an outage probability criterion. Finally, Section 5 summarizes the main results of the paper and offers some concluding remarks.

2 Average Output SNR Criterion

The average combined SNR $\overline{\gamma}_{\text{gsc}}$ at the GSC output is given by [7, Eq. (7)], [9, Eq. (16)], [17, Eq. (22)] $\overline{\gamma}_{\text{gsc}} = \left(1 + \sum_{l=L_c+1}^{L} \frac{1}{l}\right) L_c \overline{\gamma}$, where $\overline{\gamma}$ is the average SNR per symbol per diversity path. Under the i.i.d. assumption the total captured average SNR per symbol at the receiver is $\overline{\gamma}_t = L\overline{\gamma}$.

For a fixed $\overline{\gamma}_t$ and L_c the objective is to find the optimum L^* such that $\overline{\gamma}_{\text{gsc}} = \left(1 + \sum_{l=L_c+1}^{L} \frac{1}{l}\right) \frac{L_c \overline{\gamma}_t}{L}$, is maximum. This is equivalent to finding the L^* that maximizes the function

$$f(L_c, L) = \left(1 + \sum_{l=L_c+1}^{L} \frac{1}{l}\right) \frac{L_c}{L}. \tag{1}$$

Hence L^* can be obtained by differentiating $f(L_c, L)$ with respect to L. As such we need to re-write $f(L_c, L)$ in a form which does not restrict L to integer values. For this purpose, we approximate the sum in (1) by an integral, namely

$$1 + \sum_{l=L_c+1}^{L} \frac{1}{l} \approx 1 + \int_{L_c+1}^{L+1} \frac{dx}{x} = 1 + \ln\left(\frac{L+1}{L_c+1}\right). \tag{2}$$

Substituting (2) in (1), taking the derivative with respect to L, and equating the resulting expression to zero yields the transcendental equation

$$\ln\left(\frac{L_c+1}{L+1}\right) = \frac{1}{L+1}, \tag{3}$$

that L^* has to satisfy in order to maximize the average combined SNR $\overline{\gamma}_{\text{gsc}}$ at the GSC output. It can be verified that (3) is achieved for $L^* \approx L_c$ and therefore the maximum that $\overline{\gamma}_{\text{gsc}}$ can achieve is $\overline{\gamma}_t$ in agreement with the principle of "energy conservation". This is confirmed in Fig. 1 which plots the normalized average combined SNR $\overline{\gamma}_{\text{gsc}}/\overline{\gamma}_t$ versus the number of available diversity paths L. Hence, in situations where increasing the average strength of the link (rather then optimizing the BEP performance) is of primary interest, the system should not "over-spread" the transmitted signal beyond the number of paths being combined. As a byproduct, systems employing this spreading strategy are going to be more spectrally efficient in, for example, a tracking or channel estimation mode where no information data are needed to be transmitted.

For this performance criterion, because of the linear relation between $\overline{\gamma}_{\text{gsc}}$ and $\overline{\gamma}_t$, the optimum L that minimizes $\overline{\gamma}_t$ (for a fixed $\overline{\gamma}_{\text{gsc}}$ and L_c) has to minimize $f(L_c, L)$ and must therefore also be a solution of (3), i.e., it is the same $L^* \approx L_c$ that maximizes $\overline{\gamma}_{\text{gsc}}$ for a fixed $\overline{\gamma}_t$.

3 Average Probability of Error Criterion

The conditional BEP $P_b(E/\gamma)$ of several uncoded and coded modulation schemes can be upper bounded/approximated by purely exponential Chernoff-type bounds in the form

$$P_b(E/\gamma) \leq a\, e^{-b\gamma}, \tag{4}$$

where γ is the conditional SNR per bit and a and b are modulation-dependent parameters. For example, for binary coherent phase-shift-keying (BPSK) $a = 1/2$ and $b = 1$. More generally, for coherent M-PSK with Gray coding $a = (M-1)/(M \log_2(M))$ and $b = \log_2(M) \sin^2(\pi/M)$ (see [23, p. 199]), whereas for coherent M-ary quadrature modulations (QAM) with two dimensional Gray coding [24] $a = 2$ and $b = 3/(2(M-1)\log_2(M))$. Based on the bound presented in [25, Eq. (17)], it can also be shown that the conditional BEP of noncoherent BFSK with post-detection EGC can also be upper bounded by (4) with $a = 1/2$ and $b = 1/2$. Similarly, purely exponential Chernoff-type bounds/approximations have been developed for trellis coded QAM modulations (see for example [26]). We will use the bound/approximation of (4) in our analysis since its simple form leads to a generic optimization procedure for various modulation/detection schemes.

Averaging the conditional BEP (4) over the probability density function (PDF) of γ_{gsc}, yields the average BEP as

$$P_b(E) \le a\ \mathcal{M}_{\gamma_{\text{gsc}}}(-b), \tag{5}$$

where $\mathcal{M}_{\gamma_{\text{gsc}}}(\cdot)$ is the moment generating function (MGF) of γ_{gsc} obtained in closed form in [17, Eq. (12)]. Substituting [17, Eq. (12)] in (5) yields

$$P_b(E) \le \frac{a}{(1+b\overline{\gamma})^{L_c-1}\prod_{l=L_c}^{L}\left(1+\frac{b\overline{\gamma}L_c}{l}\right)}, \tag{6}$$

where $\overline{\gamma}$ is the energy per bit per diversity path.

3.1 Minimum Average Probability of Error

3.1.1 Analysis

In this section, the objective is to minimize the average BEP of (6) with respect to the number of available paths L, for a fixed number of combined paths L_c and subject to a constraint on the total captured SNR $\overline{\gamma}_t = L\overline{\gamma}$. Mathematically speaking, defining $\tilde{\gamma}_t = b\overline{\gamma}_t$, the optimum number of diversity paths is given by $L^* = L$ such that $\frac{a}{\left(1+\frac{\tilde{\gamma}_t}{L}\right)^{L_c-1}\prod_{l=L_c}^{L}\left(1+\frac{\tilde{\gamma}_t L_c}{l\ L}\right)}$, is minimum, or equivalently such that

$$\left(1+\frac{\tilde{\gamma}_t}{L}\right)^{L_c-1}\left(\frac{L_c}{L}\right)^{L-L_c+1}\frac{\prod_{l=L_c}^{L}\left(\tilde{\gamma}_t+\frac{lL}{L_c}\right)}{L!} \tag{7}$$

is maximum. Taking the natural logarithm of (7) yields the cost function

$$g(L,L_c,\tilde{\gamma}_t) = (L_c-1)\ln\left(1+\frac{\tilde{\gamma}_t}{L}\right)+(L-L_c+1)\ln\left(\frac{L_c}{L}\right)-\ln(L!)+\sum_{l=L_c}^{L}\ln\left(\tilde{\gamma}_t+\frac{lL}{L_c}\right). \tag{8}$$

To find the optimum L we need to differentiate (8) with respect to L. As such we need to re-write $g(L,L_c,\tilde{\gamma}_t)$ in a form which does not restrict L to integer values. For this purpose, borrowing some of the tricks used in [22], we first use Stirling's approximation [22, Eq. (A2)]

$$\ln(L!) \approx 0.919+\left(L+\frac{1}{2}\right)\ln L - L \tag{9}$$

in (8) yielding

$$g(L,L_c,\tilde{\gamma}_t)\approx(L_c-1)\ln\left(1+\frac{\tilde{\gamma}_t}{L}\right)+(L-L_c+1)\ln\left(\frac{L_c}{L}\right)-0.919-\left(L+\frac{1}{2}\right)\ln L+L+\sum_{l=L_c}^{L}\ln\left(\tilde{\gamma}_t+\frac{lL}{L_c}\right). \tag{10}$$

In addition, approximating the sum in (10) by an integral, namely

$$\begin{aligned}\sum_{l=L_c}^{L}\ln\left(\tilde{\gamma}_t+l\frac{L}{L_c}\right) &\approx \frac{L_c}{L}\int_{\tilde{\gamma}_t+L_c\left(\frac{L}{L_c}\right)}^{\tilde{\gamma}_t+(L+1)\frac{L}{L_c}}\ln x\ dx\\ &= \frac{L_c}{L}\left[\left(\tilde{\gamma}_t+\frac{L(L+1)}{L_c}\right)\left(\ln\left[\tilde{\gamma}_t+\frac{L(L+1)}{L_c}\right]-1\right)-(\tilde{\gamma}_t+L)(\ln(\tilde{\gamma}_t+L)-1)\right]\end{aligned} \tag{11}$$

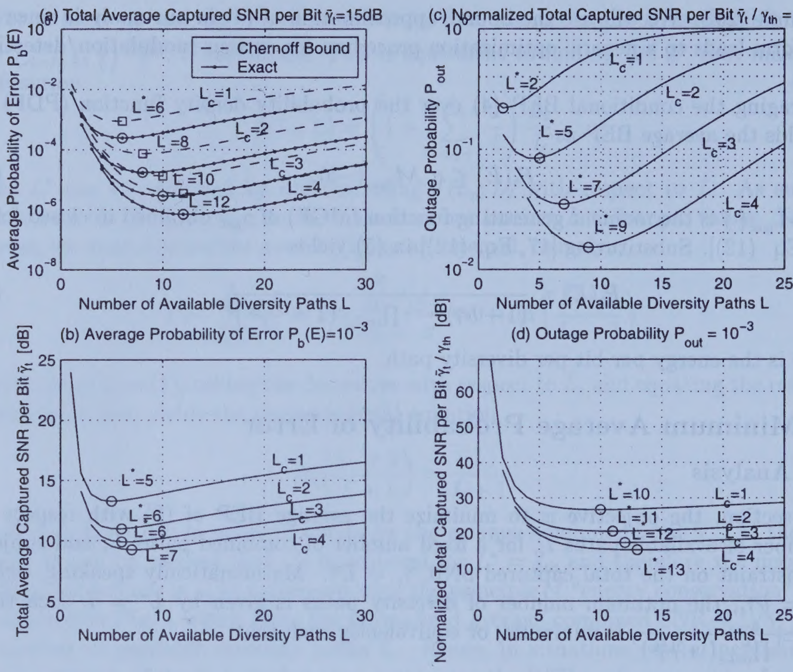


Figure 2: (a) Average BEP $P_b(E)$ of BPSK as function of the number of available paths L for various values of the combined paths L_c and a fixed total average captured SNR $\overline{\gamma}_t$. (b) Total average captured SNR $\overline{\gamma}_t$ as function of the number of available paths L for various values of the number of combined paths L_c and a fixed average BEP for BPSK. (c) Outage probability P_{out} as function of the number of available paths L for various values of the combined paths L_c and a fixed normalized total average captured SNR $\overline{\gamma}_t/\gamma_{\text{th}}$. (d) Total average captured SNR $\overline{\gamma}_t$ as function of the number of available paths L for various values of the number of combined paths L_c and a predetermined outage probability.

yields after some simplification an approximate expression for $g(L, L_c, \tilde{\gamma}_t)$ given by

$$\begin{aligned} g(L, L_c, \tilde{\gamma}_t) \;\; &\approx \;\; -1.919 + L_c - \left(1 + \frac{L_c}{L}\tilde{\gamma}_t\right) \ln\left(\tilde{\gamma}_t + L\right) - \left(2L + \frac{1}{2}\right) \ln L \\ &+ \;\; (L - L_c - 1) \ln L_c + \left(L + 1 + \frac{L_c}{L}\,\tilde{\gamma}_t\right) \ln\left(\tilde{\gamma}_t + \frac{L(L+1)}{L_c}\right), \end{aligned} \quad (12)$$

which has to be optimized with respect to L. Taking the derivative of (12) with respect to L and equating the resulting expression to zero yields the transcendental equation

$$\begin{aligned} 2 \;+\; &\frac{1}{2L} + \ln\left(\frac{L^2}{L_c}\right) + \frac{1 + \frac{L_c}{L}\tilde{\gamma}_t}{\tilde{\gamma}_t + L} - \frac{L_c}{L^2}\,\tilde{\gamma}_t \ln(\tilde{\gamma}_t + L) \\ = \;\; &\frac{(2L+1)(L + 1 + \frac{L_c}{L}\tilde{\gamma}_t)}{L_c\tilde{\gamma}_t + L(L+1)} + \left(1 - \frac{L_c}{L^2}\,\tilde{\gamma}_t\right) \ln\left(\tilde{\gamma}_t + \frac{L(L+1)}{L_c}\right) \end{aligned} \quad (13)$$

that L^* has to satisfy in order to minimize the average BEP.

3.1.2 Numerical Examples and Discussion

In this numerical examples section, we focus on the BPSK case. Fig. 2-a shows the exact (as per [17, Eq. (40)] with $\overline{\gamma}$ replaced by $\overline{\gamma}_t/L$) and the Chernoff bound (as per (6) with $a = 1/2$ and $b = 1$) on average BEP as a function of the number of available paths L for various values of L_c and for fixed $\overline{\gamma}_t$. It can be observed that the optimum solutions L^* obtained via the exact or the Chernoff bound on BEP match very well for several parameter combinations. For example, from Fig. 2-a for $\overline{\gamma}_t = 15$ dB, the L^* acquired from the bound matches exactly with the exact solution. On the other hand, on another numerical example not included in this paper, L^* is sometimes slightly off from the true optimum when $\overline{\gamma}_t = 20$ dB. However, one can see from all these plots that the average BEP curves are very smooth around the optimum. Therefore an error of one in the value of L (which is typically the value picked by the optimization of the Chernoff-bound when it is incorrect) has a negligible impact on the optimum average BEP. This validates the basis of our optimization procedure and allows us to conclude that an optimization based on the Chernoff bound can be safely used to predict the optimum spreading bandwidth for various other modulation/detection schemes.

We now propose a rule of thumb type formula to choose the optimum L for a fixed L_c and $\tilde{\gamma}_t$. For this purpose, the transcendental expression (13) for the minimum average BEP is solved numerically. The results are plotted in Fig. 3-a as stars after rounding the solution to the closest integer number towards minus infinity since this is actually a discrete optimization problem in which the integer part of L^* obtained via (13) is of interest. In Fig. 3-a, the staircase curves are the true optimum number of available paths L^* obtained via a numeric-search procedure for all possible L and using the Chernoff bound. Note the very good match obtained over the whole range of $\tilde{\gamma}_t$ between the staircase curves and the stars (after also being connected with a staircase curve). This validates the numerical optimization procedure (in particular the Stirling's and the sum/integral approximations) which can be safely used instead of the numeric-search approach. A careful look at the optimum number of paths obtained via the numerical optimization procedure (the stars in Fig. 3-a) suggests a linear relationship between L^* and $\tilde{\gamma}_t$ (on a logarithmic scale), namely $\ln L^* = \ln \alpha + \beta \ \ln \tilde{\gamma}_t$, where α and β are obtained using a least mean square fitting method and are tabulated for $L_c = 1$ to $L_c = 10$ in Table 1-a.

A close look at the values of α for various values of L_c suggests a linear relation of the form $\alpha = u + v \ L_c$, where the constants $u = 1.114$ and $v = 0.4$ are obtained again via a least mean square fitting method. Hence the proposed rule of thumb formula is given by

$$L^* \approx (1.114 + 0.4 \ L_c) \ \tilde{\gamma}_t^{0.44} \tag{14}$$

To validate this formula, it is plotted as a solid line in Fig. 3-a as a comparison with the results obtained through the Chernoff bound optimization and again in Fig. 3-b as a comparison with the results obtained via a numeric-search for all possible L and using exact average BEP expressions. The fit turns out to be a very good match with the staircase curves in both case, which allows us to conclude that the proposed formula (14) can serve as a simple and accurate reference for picking the optimum L^* for a fixed L_c and $\tilde{\gamma}_t$.

3.2 Minimum Total Average Captured SNR

We now look at the reverse optimization problem where we seek to minimize the total captured SNR $\overline{\gamma}_t = L\overline{\gamma}$ with respect to the number of available paths L for a fixed number of combined paths L_c and a required fixed average BEP. This optimization problem is solved for BPSK by fixing the average BEP in [17, Eq. (40)] and looking through a numeric-search procedure for the optimum L. Some typical resulting numerical examples are shown in Fig. 2-b. All curves show a clear minimum for $\overline{\gamma}_t$ for a given average BEP

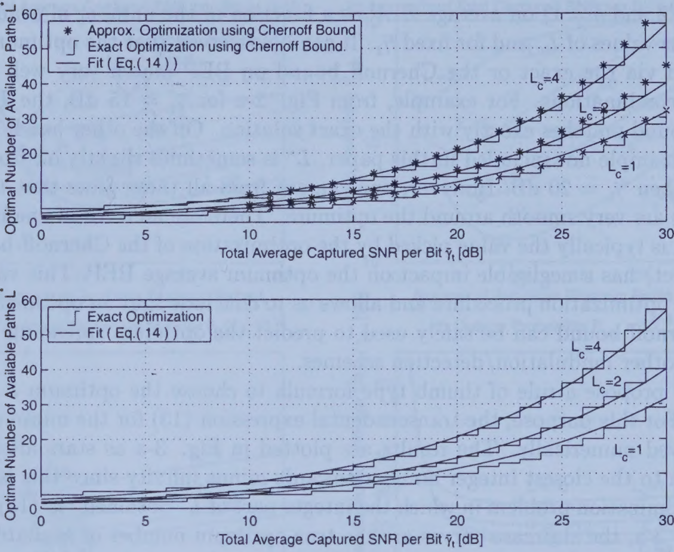


Figure 3: Dependence of the optimum number of available paths L^* as function of the total average captured energy per bit $\bar{\gamma}_t$.

and L_c. In addition the curves decrease very sharply with respect to L before reaching their minimum. Hence "under-spreading" can cost a lot in terms of transmitted energy. In contrast, these same curves tend to increase more slowly beyond the optimum L and "over-spreading", although undesirable from a spectral efficiency perspective, is less critical from a transmitted energy point of view.

4 Outage Probability Criterion

In certain communication system applications, outage probability is the most relevant performance criterion. It is defined as the probability that the instantaneous BEP exceeds a specified value or equivalently that the (instantaneous) combined SNR, γ_t, falls below a certain specified threshold γ_{th}. Similar to the previous sections, we studied the optimum spreading bandwidth for the outage probability criterion. As a numerical example, Fig. 2-c plots the outage probability (as per [17, Eq. (24)]) as a function of L for a fixed normalized total captured SNR per bit $\overline{\gamma}_t/\gamma_{\text{th}}$. As was the case for the average BEP criterion, a clear minimum for the outage probability is obtained resulting in an optimum value of L. The corresponding optimum spreading bandwidth increases with the number of combined paths. Following the same fitting procedure as for the average BEP we eventually arrive at a relation between L^* and $\overline{\gamma}_t/\gamma_{\text{th}}$ given by $\ln L^* = \ln \alpha + \beta \ \ln(\overline{\gamma}_t/\gamma_{\text{th}})$, where α and β were obtained again using a least mean square fitting method and the results are tabulated for $L_c = 1$ to $L_c = 4$ in Table 1-b.

Also, similar to the rule of thumb type of relation (14) obtained for the average BEP of BPSK, we found that a good reference for picking the optimum L^* for a fixed L_c and $\overline{\gamma}_t/\gamma_{\text{th}}$ for an outage probability criterion is given by $L^* \approx (-0.162 + 0.93 L_c)\ (\overline{\gamma}_t/\gamma_{\text{th}})^{0.89}$.

In the reverse optimization problem the total captured SNR $\overline{\gamma}_t = L\overline{\gamma}$ is minimized with respect to the number of available paths L for a fixed number of combined paths

Table 1: (a) Parameters α and β for BPSK with coherent combining. (b) Parameters α and β for the outage probability criterion.

Table	(a)		Table	(b)	
L_c	β	$\ln(\alpha)$	L_c	β	$\ln(\alpha)$
1	0.4400	0.3955	1	0.9719	-0.3147
2	0.4547	0.6437			
3	0.4545	0.8435			
4	0.4511	1.005	2	0.9234	0.4998
5	0.4467	1.1415			
6	0.4420	1.2601			
7	0.4373	1.3656	3	0.8674	1.0470
8	0.4325	1.4608			
9	0.4279	1.5479	4	0.8674	1.2335
10	0.4233	1.6282			

L_c and a required predetermined outage probability. Fig. 2-d shows a clear minimum for $\overline{\gamma}_t$ for a given average P_{out} and L_c. As in the average BEP case the curves decrease very sharply with respect to L before reaching their minimum, whereas they tend to increase more slowly beyond the optimum L. So the same remarks made in Section 3.2 for the average BEP criterion regarding "under-spreading" and "over-spreading" apply also for the outage probability criterion.

5 Conclusion

We studied the optimum spreading bandwidth of GSC receivers for various performance criteria. We offered in some instances an accurate approximate estimate of this optimum bandwidth in the form of a solution of a transcendental equation. In other cases, where the optimization is not easily tractable in an analytic fashion, we used a numeric-search procedure to find this optimum bandwidth for different performance criteria and system parameters of interest. In all cases, we adopted a least mean square fitting approach to offer some simple rule of thumb type of formulas. Validation of these formulas by comparing them with the exact results show that they can be safely used as a good reference for picking the optimum spreading bandwidth given a set of system parameters and a particular performance criterion of interest. Finally, we conclude this paper by mentioning that the study in this paper was limited to i.i.d Rayleigh diversity paths. In view of the recently derived performance results of GSC in more general fading settings (e.g., non identically distributed Rayleigh or Nakagami paths, as discussed in the introduction) the study conducted in this paper can be extended to these scenarios of practical interest. However, given the relatively complicated form of the performance results obtained, it appears that an analytical optimization procedure is not possible and one has therefore to resort to an exhaustive numeric-search procedure to have an estimate for the optimum bandwidth for a particular scenario of interest.

References

[1] F. Adachi, M. Sawahashi, and H. Suda, "Wideband DS-CDMA for next-generation mobile communications systems," *IEEE Commun. Mag.*, vol. 36, pp. 56–69, September 1998.

[2] M. Z. Win and R. A. Scholtz, "On the energy capture of ultrawide bandwidth signals in dense multipath environments," *IEEE Commun. Letters*, vol. 2, pp. 245–247, September 1998.

[3] M. Z. Win and R. A. Scholtz, "Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications," *IEEE Trans. Commun.*, vol. COM-48, pp. 679–691, April 2000.

[4] N. Kong, T. Eng, and L. B. Milstein, "A selection combining scheme for RAKE receivers," in *Proc. IEEE Int. Conf. Univ. Personal Comm. ICUPC'95, Tokyo, Japan*, pp. 426–429, November 1995.

[5] T. Eng, N. Kong, and L. B. Milstein, "Comparison of diversity combining techniques for Rayleigh-fading channels," *IEEE Trans. on Commun.*, vol. COM-44, pp. 1117–1129, September 1996. See also "Correction to "Comparison of diversity combining techniques for Rayleigh-fading channels"", *IEEE Trans. Commun.*, vol. COM-46, p. 1111, September 1998.

[6] K. J. Kim, S. Y. Kwon, E. K. Hong, and K. C. Whang, "Comments on "Comparison of diversity combining techniques for Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. COM-46, pp. 1109–1110, September 1998.

[7] N. Kong and L. B. Milstein, "Average SNR of a generalized diversity selection combining scheme," *IEEE Commun. Letters*, vol. 3, p. 5759, March 1999. See also *Proc. IEEE Int. Conf. on Commun. ICC'98, Atlanta, GA*, pp. 1556-1560, June 1998.

[8] N. Kong and L. B. Milstein, "A closed form expression for the average SNR when combining an arbitrary number of diversity branches with non-identical Rayleigh fading statistics," in *Proc. IEEE Int. Conf. on Commun. ICC'99, Vancouver, British Columbia, Canada*, pp. 1864–1868, June 1999.

[9] M. Z. Win and J. H. Winters, "Analysis of hybrid selection/maximal-ratio combining in Rayleigh fading," *IEEE Trans. Commun.*, vol. COM-47, pp. 1773–1776, December 1999. See also *Proceedings of IEEE International Conference on Communications (ICC'99)*, pp. 6-10, Vancouver, British Columbia, Canada, June 1999.

[10] M. Z. Win and J. H. Winters, "Analysis of hybrid selection/maximal-ratio combining of diversity branches with unequal SNR in Rayleigh fading," in *IEEE Veh. Technol. Conf. (VTC'99), Houston, TX*, pp. 215–220, May 1999.

[11] M. Z. Win and Z. A. Kostić, "Virtual path analysis of selective Rake receiver in dense multipath channels," *IEEE Commun. Letters*, vol. 3, pp. 308–310, November 1999. See also *Proceedings of 8th Communication Theory Mini-Conference (CTMC VIII) in conjunction with IEEE International Conference on Communications (ICC'99)*, pp. 78-82, Vancouver, British Columbia, Canada, June 1999.

[12] M. Z. Win and Z. A. Kostić, "Impact of spreading bandwidth on Rake reception in dense multipath channels," *IEEE J. Select. Areas Commun.*, vol. SAC-17, pp. 1794–1806, October 1999. See also *Proceedings of 8th Communication Theory Mini-Conference (CTMC VIII) in conjunction with IEEE International Conference on Communications (ICC'99)*, pp. 78-82, Vancouver, British Columbia, Canada, June 1999.

[13] M. Z. Win, R. K. Mallik, G. Chrisikos, and J. H. Winters, "Canonical expressions for the error probability performance for M-ary modulation with hybrid selection/maximal-ratio combining in Rayleigh fading," in *IEEE Wireless Commun. and Networking Conf. (WCNC'99), New Orleans, LA*, pp. 266–270, September 1999.

[14] M. Z. Win, G. Chrisikos, and N. R. Sollenberger, "Impact of spreading bandwidth and diversity order on the error probability performance of Rake reception in dense multipath channels," in *IEEE Wireless Commun. and Networking Conf. (WCNC'99), New Orleans, LA*, pp. 1558–1562, September 1999.

[15] M. Z. Win, G. Chrisikos, and J. H. Winters, "Error probability performance for M-ary modulation in Rayleigh fading using hybrid selection/maximal-ratio combining," in *Proc. IEEE Mil. Commun. Conf. (MILCOM'99)*, November 1999.

[16] M. Z. Win and J. H. Winters, "Exact error probability expressions for H-S/MRC in Rayleigh fading: A virtual branch technique," in *IEEE Global Commun. Conf. (GLOBECOM'99)*, pp. 537–541, December 1999.

[17] M. -S. Alouini and M. K. Simon, "An MGF-based performance analysis of generalized selective combining over Rayleigh fading channels," *IEEE Trans. Commun.*, vol. COM-48, pp. 401–415, March 2000. See also *Proceedings of 8th Communication Theory Mini-Conference (CTMC VIII) in conjunction with IEEE International Conference on Communications (ICC'99)*, pp. 110-114, Vancouver, British Columbia, Canada, June 1999.

[18] M. -S. Alouini and M. K. Simon, "Performance of coherent receivers with hybrid SC/MRC over Nakagami-m fading channels," *IEEE Trans. Veh. Technol.*, vol. VT-48, pp. 1155–1164, July 1999.

[19] M. -S. Alouini and M. K. Simon, "Application of the Dirichlet transformation to the performance evaluation of generalized selection combining over Nakagami-m fading channels," *Int. J. of Commun. and Networks*, vol. 1, pp. 5–13, March 1999. See also *Proc. of the IEEE Vehicular Technology Conference (VTC'99-Fall)*, pp. 953-957, Amsterdam, The Netherlands, September 1999.

[20] J. A. Ritcey and M. Azizoğlu, "Performance analysis of generalized selection combining with switching constraints," *IEEE Commun. Letters*, vol. 4, pp. 152–154, May 2000. See also *Proc. Conf. on Information Sceinces and Systems (CISS'2000)*, Princeton, NJ, pp. WA8a-1-WA8a-6, March 2000.

[21] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering.* New York, Wiley, 1965.

[22] C. Leung, "Optimized selection diversity for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. COM-30, pp. 554–557, March 1982.

[23] M. K. Simon and M. -S. Alouini, "Digital communications over generalized fading channels: A unified approach to performance analysis." New York, NY: John Wiley & Sons, to be published in July 2000.

[24] G. J. Foschini and J. Salz, "Digital communications over fading radio channels," *Bell Syst. Tech. J.*, pp. 429–456, February 1983.

[25] M. K. Simon and M. -S. Alouini, "Exponential-type bounds on the generalized Marcum Q-function with application to error probability analysis over fading channels," *IEEE Trans. Commun.*, vol. COM-48, pp. 359–366, March 2000.

[26] K. J. Hole, H. Holm, and G. E. Øien, "Adaptive multidimensional coded modulation over flat fading channels," *IEEE J. Select. Areas Commun.*, vol. SAC-18, pp. 1153–1158, July 2000.

Performance Analysis of Hierarchical Coded 16-QAM Systems*

J.-H. Lim and S. B. Gelfand
School of Electrical and Computer Engineering
Purdue University
W. Lafayette, IN, 47907-1285
jonghan@ecn.purdue.edu and gelfand@ecn.purdue.edu

Abstract

Analytical performance bounds are presented for a hierarchical coded modulation system which uses two convolutional encoders (and two Viterbi decoders) with a hierarchical 16-QAM constellation. This system is compatible with an existing QPSK based coded system and/or is suitable for variable error rate protection. We consider four different schemes (two decoding schemes, multi-stage decoding(MSD) and independent decoding(ID), and two labeling schemes, Gray labeling(GL) and non-Gray labeling(NGL)) in the AWGN channel and show that the performance of the simple GL with ID scheme is nearly optimum. We also analyze the GL with ID scheme in the Rayleigh fading channel. Finally, we compare our bound with the expurgated bound for bit interleaved coded modulation, and also bounds for multilevel coded modulation with staged decoding, specialized to the hierarchical coded system.

1 Introduction

The hierarchical modulation scheme in Fig. 1 [1] uses two convolutional encoders with a hierarchical 16-QAM constellation (see Fig. 2). To maintain the backward compatibility with an existing QPSK modulation, the output of convolutional encoder-I is fed to the two LSB bits and that of the convolutional encoder-II is fed to the two MSB bits. Bit interleaving is used to improve performance in fading channels, and even in the AWGN channel since it can reduce the error propagation from level-I bits to level-II bits in the case of multi-stage decoding (only bit interleaver-II is needed in the AWGN channel with multi-stage decoding). Since hierarchical 16-QAM preserves the appearance of the QPSK signal, it can operate with an existing QPSK receiver (which decodes only level-I bits). In addition to the backward compatibility, it can also be used as unequal error protection, that is, to transmit information in the DVB-T system [2] with higher reliability according to its importance.

Multilevel coded modulation [3], [4] is a similar scheme, which uses multiple encoders and staged decoding. However, its signal mapper is usually that of TCM, which is not compatible with an existing QPSK system. Pragmatic TCM [5] is also a related

---

*Research supported by Thomson Consumer Electronics, and the National Science Foundation under Grant NCR-9706372.

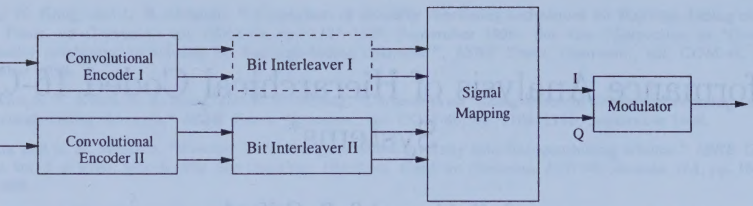
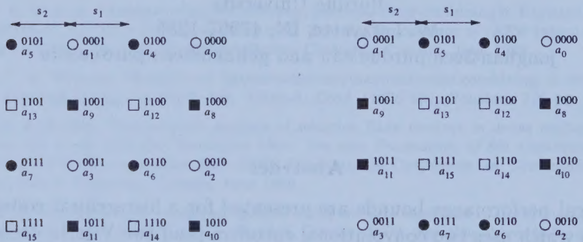

Convolutional Encoder I
Bit Interleaver I
Convolutional Encoder II
Bit Interleaver II
Signal Mapping
I
Q
Modulator

Figure 1: QPSK Compatible Coded Modulation System.


s_2
s_1
0101 a_5
0001 a_1
0100 a_4
0000 a_0
1101 a_{13}
1001 a_9
1100 a_{12}
1000 a_8
0111 a_7
0011 a_3
0110 a_6
0010 a_2
1111 a_{15}
1011 a_{11}
1110 a_{14}
1010 a_{10}
s_2
s_1
0001 a_1
0101 a_5
0100 a_4
0000 a_0
1001 a_9
1101 a_{13}
1100 a_{12}
1000 a_8
1011 a_{11}
1111 a_{15}
1110 a_{14}
1010 a_{10}
0011 a_3
0111 a_7
0110 a_6
0010 a_2

(a) (b)

Figure 2: 16-QAM Mapper for Hierarchical Coding. (a) Non-Gray Labeling. (b) Gray Labeling.

scheme. Its purpose is to use a single basic coder and decoder (*de facto* standard (133, 171) convolutional code) in multiple coding applications. The pragmatic trellis-coded 16-QAM uses two separate (133, 171) convolutional encoders and decoders (bandwidth efficiency = 2 b/s/Hz) and its performance is slightly better than that of Ungerboek's 64 state 8-PSK TCM. But its signal mapper maps each code to I and Q separately, and is also not compatible with an existing QPSK based system.

In this paper, we present analytical bounds on the bit error rate (BER) of the QPSK compatible hierarchically coded modulation system in Fig. 1. In Section 2, we present analytical upper bounds for four different schemes (two decoding schemes, multi-stage decoding(MSD) and independent decoding(ID), and two labeling schemes, Gray labeling(GL) and non-Gray labeling(NGL)) in AWGN channel and show that the performance of the simple Gray labeling with independent decoding scheme is nearly optimum. We extend the analysis of the system in Rayleigh channel by focusing on Gray labeling with independent decoding scheme in Section 3 (when perfect channel state information (CSI) is available). Simulation results are given for comparison with the bounds in Section 4, and in Section 5 we compare our bound with the expurgated bound for bit interleaved coded modulation (BICM) [7], and also bounds for multilevel coded modulation (MCM) [8] with staged decoding, specialized to the hierarchical coded system.

2 Upper Bound in AWGN Channel

When BPSK or QPSK is used, the average bit error probability, P_b of (n,k) rate coded system is bounded by

$$P_b \;\leq\; \frac{1}{k}\sum_{d=d_{\text{free}}}^{\infty} v(d)P_2(d) \tag{1}$$

where $P_2(d)$ is the pairwise error probability of two paths that differ by d codebits. $v(d)$ of a linear code can be calculated by the transfer function method. Another useful probability is the average *codebit* error probability, P_c. It is bounded by

$$P_c \le \frac{1}{n} \sum_{d=d_{\text{free}}}^{\infty} v_c(d) P_2(d).$$

From the definition of transfer function $(T(D,N))$, $v_c(d)$ can be obtained from

$$D \frac{\partial T(D,N)}{\partial D}\bigg|_{N=1} = \sum_{d=d_{\text{free}}}^{\infty} v_c(d) D^d.$$

The exact estimate of pairwise error probability, $P_2(d) \equiv P(\mathbf{c} \to \mathbf{c} + \mathbf{e})$ is crucial to the calculation of tight upper bounds.

2.1 Pairwise Error Probability of Level-I

Under the assumption of ideal interleaving of level-II bits, BER's of level-I bits for Gray labeling and non-Gray labeling is the same, so we do not distinguish between these labelings in this subsection.

First, we will consider possible branch metrics for level-I decoding (refer to Fig. 2). Without the information of level-II bits, the negative log-likelihood is given by

$$M_1(y, \text{two LSB} = (0,0)) = -\log\left(\sum_{k\in\{0,4,8,12\}} \exp\left(\frac{-|y-a_k|^2}{2\sigma^2}\right)\right). \tag{2}$$

This metric function is optimal in the sense of maximum likelihood estimation when there is no available information about level-II bits. Unfortunately, (2) involves complex computations in that we have to know σ^2 and calculate log and exp functions. A less complex branch metric is

$$M_2(y, \text{two LSB} = (0,0)) = \min_{k\in\{0,4,8,12\}} \{|y-a_k|^2\}. \tag{3}$$

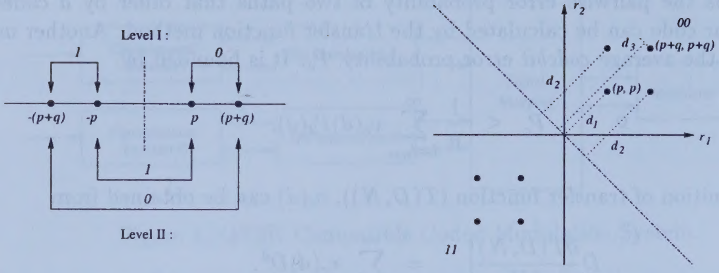
It is clear that (3) is an approximation of (2) when SNR is high. Another possible branch metric (using innermost point only) is

$$M_3(y, \text{two LSB} = (0,0)) = |y - a_{12}|^2. \tag{4}$$

We will focus on branch metric M_3 in the analysis since it is the most tractable. It turns out that the above three branch metrics do not make a large performance difference in the simulations.

Fig. 3 shows the locations of level-I coded signals. Define the n-dimensional random vector $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_n)$, $\delta_k \in \{0,1\}$, such that when the code sequence is '000 $\cdots$ 00', the signal is located at $p\mathbf{1} + q\boldsymbol{\delta}$ (the exact point is determined by the level-II code sequence). Moreover, the decision boundary is obviously the $(n-1)$-dimensional hyperplane represented by $\sum_{k=1}^{n} x_k = 0$ when branch metric M_3 is used in the decoding. Now the distance from the point $p\mathbf{1} + q\boldsymbol{\delta}$ to the hyper-plane $\sum_{k=1}^{n} x_k = 0$ is

$$d = \frac{pn + q(\mathbf{1} \bullet \boldsymbol{\delta})}{\sqrt{n}}$$



(a) (b)

Figure 3: Locations of Level-I Code in One or Two Dimensional Space (Gray Labeling). (a) One Dimensional Space. (b) Two Dimensional Space.

where $\bullet$ denotes vector inner product. So, when the δ_k's are i.i.d. with $P_I(\delta_k = 0) = P_I(\delta_k = 1) = 1/2$ (this is valid under ideal interleaving of level-II bits),

$$P_I(\mathbf{c} \to \mathbf{c} + \mathbf{e}) \;=\; \mathrm{E}_{\delta}\left\{ Q\left(\frac{pn + q(\mathbf{1} \bullet \delta)}{\sigma\sqrt{n}} \right) \right\} = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{1}{2}\right)^{n} Q\left(\frac{pn + qk}{\sigma\sqrt{n}} \right)$$

when $H(\mathbf{e}) = n$ ($H(\cdot)$ denotes Hamming distance).

2.2 Pairwise Error Probability of Level-II

2.2.1 Independent Decoding (ID)

We assume the following metric function in independent decoding of level-II bits :

$$M(y, \text{two MSB} = (0,0)) \;=\; \min_{k \in \{0,1,2,3\}} \{|y - a_k|^2\}$$

Using similar reasoning as Section 2.1, we can show that

$$P_{II,\mathrm{ID,GL}}(\mathbf{c} \to \mathbf{c} + \mathbf{e}) \;\leq\; \sum_{k=0}^{n} \frac{1}{2^k} \binom{n}{k} Q\left(\frac{qn/2 + 2pk}{\sigma\sqrt{n}} \right)$$

for Gray labeling and

$$P_{II,\mathrm{ID,NGL}}(\mathbf{c} \to \mathbf{c} + \mathbf{e}) \leq \sum_{k=0}^{n} \sum_{l=0}^{k} \binom{n}{k} \left(\frac{1}{2}\right)^{k} \binom{k}{l} \times Q\left(\frac{\sqrt{nq^2 + (2p+q)^2 k - (4pq + 2q^2)(k - 2l)}}{2\sigma} \right)$$

for non-Gray labeling when $H(\mathbf{e}) = n$.

2.2.2 Multi-Stage Decoding (MSD) on Non-Gray Labeling

A multi-stage decoder is shown in Fig. 4. In this scheme, decoding of level-II bits is dependent on the output of level-I re-encoder, i.e., the branch metric for the two MSB

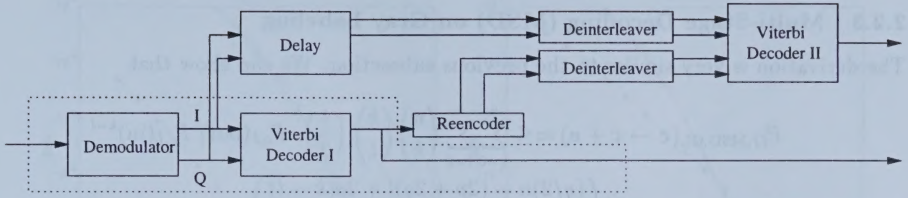
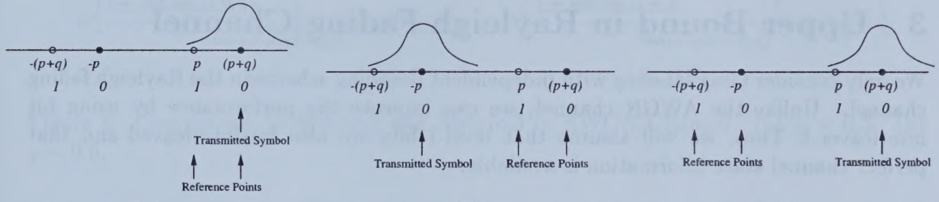

Delay
Deinterleaver
Deinterleaver
Viterbi Decoder II
I
Demodulator
Q
Viterbi Decoder I
Reencoder

Figure 4: Multi-Stage Decoding.


-(p+q)
-p
p
(p+q)
1
0
1
0
Transmitted Symbol
Reference Points
-(p+q)
-p
p
(p+q)
1
0
1
0
Transmitted Symbol
Reference Points
-(p+q)
-p
p
(p+q)
1
0
1
0
Reference Points
Transmitted Symbol

(a) (b) (c)

Figure 5: Various Situations for Level-2 Decoding, Non-Gray Labeling. (a) Correct Decision for Level-I bit. (b) Incorrect Decision for Level-I bit (Bad). (c) Incorrect Decision for Level-I bit (Good).

is dependent on the corresponding two LSB. For example, the branch metric for the two MSB being 00 is

$$\begin{aligned} M(y, \text{two MSB} = (0,0) \,|\, \text{re-encoded two LSB} = 00) &= |y - a_0|^2 \\ M(y, \text{two MSB} = (0,0) \,|\, \text{re-encoded two LSB} = 01) &= |y - a_1|^2 \\ &\cdots \end{aligned}$$

Fig. 5 shows various situations which occur in the level-II decoding. Situation (b) occurs when the transmitted symbol is an inner point (that is, closer to the origin) whereas situation (c) occurs when it is an outer point. Let $P_{c,I}(\text{in})$ ($P_{c,I}(\text{out})$) be the codebit error probability of level-I when the current symbol is an inner point (outer point). Then, the probability of situation (b) is $P_{c,I}(\text{in})$ whereas that of situation (c) is $P_{c,I}(\text{out})$. In fact,

$$\begin{aligned} P_{c,I}(\text{in}) &\leq \frac{1}{\hat{n}} \sum_{d=d_{\text{free}}}^{\infty} v_{c,I}(d) P_{2,\text{in}}(d) = \frac{1}{\hat{n}} \sum_{d=d_{\text{free}}}^{\infty} v_{c,I}(d) \sum_{k=0}^{d-1} \binom{d-1}{k} \left(\frac{1}{2}\right)^{d-1} Q\left(\frac{pd+qk}{\sigma\sqrt{d}}\right) \\ P_{c,I}(\text{out}) &\leq \frac{1}{\hat{n}} \sum_{d=d_{\text{free}}}^{\infty} v_{c,I}(d) P_{2,\text{out}}(d) = \frac{1}{\hat{n}} \sum_{d=d_{\text{free}}}^{\infty} v_{c,I}(d) \sum_{k=1}^{d} \binom{d-1}{k-1} \left(\frac{1}{2}\right)^{d-1} Q\left(\frac{pd+qk}{\sigma\sqrt{d}}\right) \end{aligned}$$

and $P_{c,I} = 0.5P_{c,I}(\text{in}) + 0.5P_{c,I}(\text{out})$. By assuming that the error status α_k of decoded/re-encoded level-I codebit is independent of the noise n_k (the dependence of α_k upon n_k becomes less as the code becomes more powerful),

$$\begin{aligned} P_{II,\text{MSD,NGL}}(\mathbf{c} \rightarrow \mathbf{c} + \mathbf{e}) \approx\leq \sum_{k=0}^{n} \sum_{l=0}^{k} \binom{n}{k} \binom{k}{l} \left(\frac{1}{2}\right)^{k} P_{c,I}(\text{in})^{l} P_{c,I}(\text{out})^{k-l} \\ \times Q\left(\frac{(q/2)n - (2p+q)l + (2p+q)(k-l)}{\sigma\sqrt{n}}\right). \end{aligned}$$

2.2.3 Multi-Stage Decoding (MSD) on Gray Labeling

The derivation is very similar to the previous subsection. We can show that

$$P_{II,\mathrm{MSD,GL}}(\mathbf{c} \to \mathbf{c} + \mathbf{e}) \approx\leq \sum_{k=0}^{n} \sum_{l=0}^{k} \binom{n}{k} \binom{k}{l} \left(\frac{1}{2}\right)^{k} P_{c,I}(\mathrm{out})^{l} P_{c,I}(\mathrm{in})^{k-l}$$
$$\times Q\left(\frac{(q/2)n - (2p+2q)l + 2p(k-l)}{\sigma\sqrt{n}}\right).$$

3 Upper Bound in Rayleigh Fading Channel

We only consider Gray labeling with independent decoding scheme in the Rayleigh fading channel. Unlike the AWGN channel, we can improve the performance by using bit interleaver-I. Thus, we will assume that level-I bits are also bit-interleaved and that perfect channel state information is available.

3.1 Pairwise Error Probability of Level-I

We will assume that branch metric M_3 is used for level-I decoding (similar to the case of AWGN channel). By a lengthy derivation, we can show that

$$P_I(\mathbf{c} \to \mathbf{c} + \mathbf{e}) \leq \frac{1}{2^n} F_n\left(\frac{p}{\sigma}\right) + \frac{1}{2^n} F_n\left(\frac{p+q}{\sigma}\right) + \sum_{k=1}^{n-1} \frac{1}{2^n} \binom{n}{k} \sum_{l=0}^{N-1} F_n\left(\frac{pN+ql}{\sigma N}\right) C_N(n,k,l)$$

for all positive integers N where

$$C_N(n,k,l) = I\left(\frac{l+1}{N}; k, n-k\right) - I\left(\frac{l}{N}; k, n-k\right)$$
$$F_n(x) = \int_0^{\infty} Q\left(x\sqrt{z}\right) \exp(-z) \frac{z^{n-1}}{(n-1)!}\, dz.$$

$I(x;p,q)$ is the incomplete beta function which is defined as $\int_0^x t^{p-1}(1-t)^{q-1}\,dt / \int_0^1 t^{p-1}(1-t)^{q-1}\,dt$. Using a similar procedure as [6], we can calculate $F_n(x)$ in an explicit form

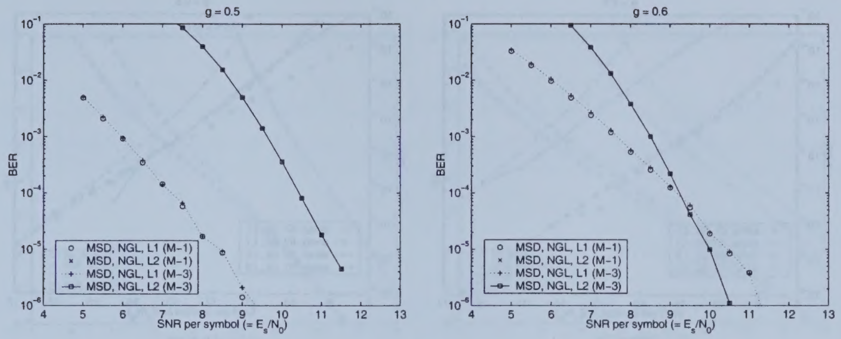
$$F_n(x) = \left[\frac{1}{2(2+x^2)}\right]^n \sum_{j=1}^{n} \binom{2n-j-1}{n-1} \frac{2^j}{\left[1+\sqrt{x^2/(2+x^2)}\right]^j}.$$

3.2 Pairwise Error Probability of Level-II

When Gray labeling and independent decoding is used,

$$P_{II,\mathrm{ID,GL}}(\mathbf{c} \to \mathbf{c} + \mathbf{e}) \leq F_n\left(\frac{q}{2\sigma}\right) + \frac{1}{2^n} F_n\left(\frac{q+4p}{2\sigma}\right)$$
$$+ \sum_{k=1}^{n-1} \frac{1}{2^k} \binom{n}{k} \sum_{l=0}^{N-1} F_n\left(\frac{qN+4pl}{2\sigma N}\right) C_N(n,k,l).$$

for all positive integers N.

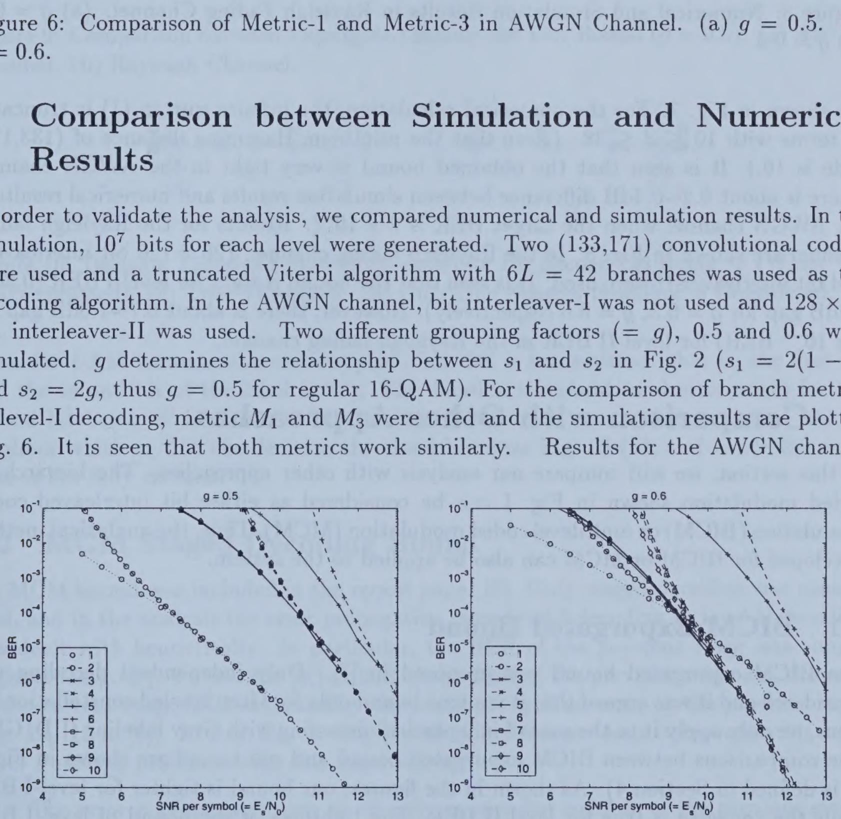


(a) (b)

Figure 6: Comparison of Metric-1 and Metric-3 in AWGN Channel. (a) $g = 0.5$. (b) $g = 0.6$.

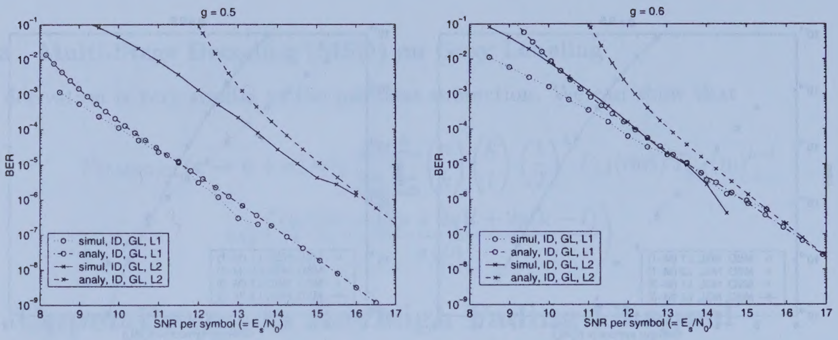
4 Comparison between Simulation and Numerical Results

In order to validate the analysis, we compared numerical and simulation results. In the simulation, 10^7 bits for each level were generated. Two (133,171) convolutional coders were used and a truncated Viterbi algorithm with $6L = 42$ branches was used as the decoding algorithm. In the AWGN channel, bit interleaver-I was not used and 128×64 bit interleaver-II was used. Two different grouping factors (= g), 0.5 and 0.6 were simulated. g determines the relationship between s_1 and s_2 in Fig. 2 ($s_1 = 2(1-g)$ and $s_2 = 2g$, thus $g = 0.5$ for regular 16-QAM). For the comparison of branch metrics in level-I decoding, metric M_1 and M_3 were tried and the simulation results are plotted Fig. 6. It is seen that both metrics work similarly. Results for the AWGN channel



(a) (b)

Figure 7: Numerical and Simulation Results in AWGN Channel. (a) $g = 0.5$. (b) $g = 0.6$. 1-[simul, L1], 2-[analy, L1], 3-[simul,MSD, NGL, L2], 4-[analy,MSD, NGL, L2], 5-[simul,ID, NGL, L2], 6-[analy,ID, NGL, L2], 7-[simul,MSD, GL, L2], 8-[analy,MSD, GL, L2], 9-[simul,ID, GL, L2], 10-[analy,ID, GL, L2].



(a) (b)

Figure 8: Numerical and Simulation Results in Rayleigh Fading Channel. (a) $g = 0.5$. (b) $g = 0.6$.

are shown in Fig. 7. For the numerical calculation, the infinite sum in (1) is truncated to terms with $10 \leq d \leq 38$. (Note that the minimum Hamming distance of (133,171) code is 10.) It is seen that the obtained bound is very tight in the AWGN channel. There is about 0.2~0.4dB difference between simulation results and numerical results in the AWGN channel when the target BER is 2×10^{-4}. Results for the Rayleigh fading channel are shown in Fig. 8. In the Rayleigh fading channel, 126×126 bit interleaver-I and bit interleaver-II were used. It is seen that the bound is tight for level-I BER (0.3dB, 0.4dB gap for $g = 0.5$, $g = 0.6$ respectively). However, there is about 0.7~1.0dB gap (at 2×10^{-4} BER) for level-II BER in the Rayleigh fading channel.

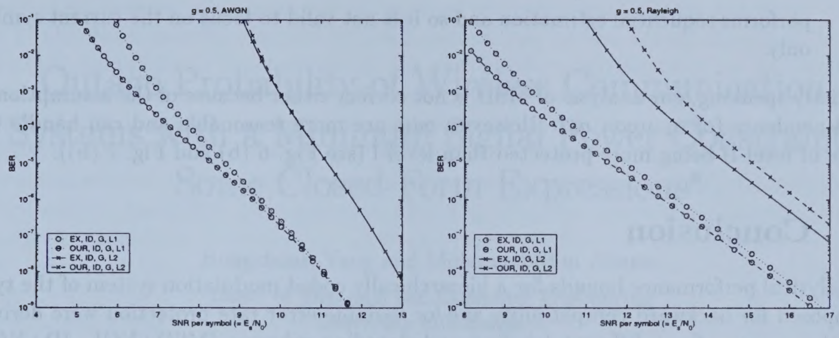
5 Comparison with Other Approaches

In this section, we will compare our analysis with other approaches. The hierarchical coded modulation shown in Fig. 1 can be considered as either bit interleaved coded modulation (BICM) or multilevel coded modulation (MCM). Thus, the analytical method developed for BICM or MCM can also be applied to the system.

5.1 BICM Expurgated Bound

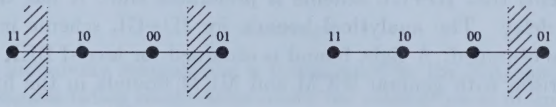
The BICM expurgated bound was proposed in [7]. Only independent decoding was considered and it was argued that it is a true bound only for Gray labeled constellation [7]. Thus, we only apply it to the case of independent decoding with Gray labeling ([ID, GL]). The comparisons between BICM expurgated bound and our bound are shown in Fig. 9 (g is defined in Section 4). As shown in the figures, our bound is tighter for level-I BER while the converse is true for level-II BER. The tightness of the bound of level-I BER shows that it is really advantageous to use branch metric M_3 in the analysis of level-I BER. (It is clear that the upper bound of metric M_3 receiver is an upper bound of the *best* receiver.)

However, the looseness of the bound of level-II BER suggests that the BICM expurgated bound may not be an upper bound even for Gray labeling. In fact, we believe that the consideration of only one unique neighbor is not sufficient for the upper bounding.



(a) (b)

Figure 9: Comparison between Expurgated Bound and Our Bound ($g = 0.5$). (a) AWGN Channel. (b) Rayleigh Channel.



(a) (b)

Figure 10: Decision Threshold (and Error Zone) for the Second Bit Being '0' (When '00' Is Transmitted). (a) True Threshold. (b) Only One Nearest Neighbor is Considered.

Consider 4-AM transmission and suppose that '00' is transmitted (that is, the first bit and the second bit of the label is '0'). The decision threshold (and error zone) for the second bit being '0' is shown in Fig. 10 (a). However, if we consider only one unique neighbor which is '01', the decision threshold becomes Fig. 10 (b), and hence the error zone is not fully covered.

5.2 MCM Staged Decoding Bound

An MCM bound was included in the recent paper [8]. Only staged decoding was considered, and in the analysis the error propagation from level-I decoding to level-II decoding was dealt with heuristically. In particular, the BER of the previous stage was simply neglected (e.g. Eq. 9 and 12 in [8]) or half of the BER of the previous stage was summed to the current stage BER obtained by the assumption of no error at the previous stage (e.g. Eq. 15 in [8]). In fact, Eq. 15 in [8] used half of the previous *information* bit error ratio. Potential problems with this approach are :

- The error is propagated due to wrongly re-encoded *codebits* of the previous stage. Thus, it is not proper to consider *information* bit error of the previous stage, especially, when the code rates of current stage and previous stage are different.
- The choice of factor 1/2 (=half) is not reasonable if the current stage is coded. If the current stage is uncoded, it may be justified by the argument of "good" situation and "bad" situation shown in Section 2.2.2. That is, we can approximate that the "bad" situation always results in an incorrect decision and the "good" situation always results in a correct decision. However, if it is coded, the decoder

performs sequential estimation and so it is not valid to focus on the current symbol only.

Strictly speaking, our analysis of MSD is not correct either because of the assumption of independence (of α_k upon n_k). However, ours are more reasonable and can handle the case of level-II being more protected than level-I (see Fig. 6 (b) and Fig. 7 (b)).

6 Conclusion

Analytical performance bounds for a hierarchically coded modulation system of the type proposed for backward compatibility and/or variable error rate protection were derived in this paper. Four different labeling and decoding schemes (MSD+NGL, ID+NGL, MSD+GL and ID+GL) were considered and compared in the AWGN channel. The obtained bounds are tight in the AWGN channel (at most 0.4dB difference at 2×10^{-4} BER). Also the performance of MSD+NGL, MSD+GL and ID+GL schemes are nearly the same. It seems that ID+GL scheme is preferable since it has the least decoding complexity and delay. The analytical bounds for ID+GL scheme in Rayleigh fading channel were also obtained. A tight bound is obtained for level-I BER but not for level-II BER. Comparison with general BICM and MCM bounds in the literature was also carried out.

References

[1] K. Ramaswamy and P. Knutson. Improved modulation and coding for satellite systems. Technical report, Thomson Consumer Electronics, 1997.

[2] ETSI Secretariat. Digital video broadcasting (DVB) : Framing structure, channel coding and modulation for digital terrestrial television. Technical Report ETS EN 300 744, European Telecommunications Standards Institute, 1999.

[3] A. R. Calderbank. Multilevel codes and multistage decoding. *IEEE Transactions on Communications*, COM-37:222–229, March 1989.

[4] Ezio Biglieri. Parallel demodulation of multidimensional signals. *IEEE Transactions on Communications*, COM-40:1581–1587, October 1992.

[5] A. J. Viterbi, J. K. Wolf, E. Zehavi, and R. Padovani. A pragmatic approach to trellis-coded modulation. *IEEE Communications Magazine*, pages 11–19, July 1989.

[6] S. Ben Slimane and T. Le-Ngoc. Tighter bounds on the error probability of coded modulation schemes in rayleigh fading channels. *IEEE Transactions on Vehicular Technology*, Vol. 44:121–129, February 1995.

[7] Giuseppe Caire, Giorgio Taricco, and Ezio Biglieri. Bit-interleaved coded modulation. *IEEE Transactions on Information Theory*, IT-44:927–946, May 1998.

[8] R. H. Morelos-Zaragoza, M. P. C. Fossorier, S. Lin, and H. Imai. Multilevel coded modulation for unequal error protection and multistage decoding - part I : Symmetric constellations. *IEEE Transactions on Communications*, 48:204–212, February 2000.

Outage Probability of Wireless Communication Systems with a Minimum Signal Power Constraint: Some Closed-Form Expressions*

Hongchuan Yang and Mohamed-Slim Alouini

Dept. of Electrical and Computer Engineering
University of Minnesota
Minneapolis, MN 55455, USA.
E-mail: <hyang,alouini@ece.umn.edu>

Abstract

Outage probability computation is investigated for the Nakagami/Rician and Rician/Nakagami fading combination under the assumption of independent identical distributed (IID) interferes. We show that with the consideration of both a minimum carrier-to-interference ratio and a minimum carrier-to-noise ratio, closed-form expressions can still be obtained in these two particular scenarios if the Nakagami parameter is restricted to integer values. These expressions are expressed in terms of the Marcum Q-function and modified Bessel functions, and therefore readily allow numerical evaluation for various cases of practical interest. When applicable, the results obtained as special cases of the offered closed-form expressions have been compared to those previously reported in the literature. In addition, for both fading combinations under consideration, numerical examples have been provided and discussed.

1 Introduction

In wireless communications systems subject to co-channel interference, outage probability is one of the most important criterion for the performance evaluation. Outage occurs when either the carrier-to-interference ratio (CIR) or the carrier-to-noise ratio (CNR) is less than their respective predetermined threshold values. This performance criterion has therefore been thoroughly studied by many researchers [1-8]. In all these publications special attention has been paid to the impact of the fading statistics of the desired and interfering users on the system outage probability. In general, when Rician fading is included in the model, closed-form solution of the problem is more difficult to obtain, especially when both CIR and CNR constraints are considered. In this paper, we consider two special fading combinations: Nakagami/Rician[1] and Rician/Nakagami, for which closed-form expressions for outage probability is given in terms of the Marcum Q-function and modified Bessel functions.

---

[*] This research is supported in part by NSF Grant No. CCR-9983462.

[1] With Nakagami/Rician, we mean that the desired signal is Nakagami faded whereas the interferers are subject to Rician fading.

In micro-cell mobile communication systems, a typical scenario consists of a direct line of sight path for the desired user accompanied by multiple diffused interferers. This scenario can be appropriately modeled with the Rician/Rayleigh combination. Yao and Sheikh [2] derived closed-form solutions for this fading combination under the assumption that all the interferers must have distinct average fading power. Lin *et al.* [8] addressed the more general Rician/Nakagami combination. In particular, they offered a simple closed-form expression for the interference limited case and an infinite series solution for the more general case where both CIR and CNR constraints are considered. In this work, we complement these two contributions by deriving a closed-form expression of outage probability for the Rician/Rayleigh scenario with the assumption of IID interferers. We then extend our analysis to cover the more general Rician/Nakagami scenario yielding an alternative closed-form expression for the infinite series representation provided in [8] for the same fading combination.

The unfavorable Rayleigh/Rician scenario applies when the spectrum is shared by competing micro-cellular communication services or by microwave users with interferers/jammers in the same band, as it is for example the case in uncoordinated ad-hoc wireless systems. Haug and Ucci [4] were the first to study this scenario. In particular they derived a closed-form outage probability expression for the single user/single interferer case. Their result was generalized by Tjhung *et al.* [5, 6] who provided a closed-form solution for the more general Rician/Rician scenario which includes the Rayleigh/Rician combination as a special case. But, both of these contributions only addressed the interference limited case. In this paper, we provide a closed-form expression for the outage probability of Rayleigh /Rician scenario where the interferers are assumed to be independent identically distributed (IID) and with the consideration of both minimum CIR and minimum signal strength requirements. We also extend this expression to the more general Nakagami/Rician combination generalizing the expression provided in [7] for the same fading scenario but for the interference limited case.

The remainder of this paper is organized as follows. Section 2 contains the general formulation of outage probability problem under consideration. The Rician/IID Nakagami fading scenario is discussed in Section 3, while Section 4 deals with the Nakagami/IID Rician case. In both sections, the special cases that use Rayleigh fading instead of Nakagami fading are first investigated, then the results are generalized to Nakagami cases. The generic closed-form results for outage probability are derived and special cases of interest are investigated, with comparison to those results previously reported in the literature whenever applicable. Also, numerical results and relevant discussion for both cases are provided. Note that our analysis limits its scope to the effect of multi-path fading and does not take into account the impact of log-normal shadowing. This can be done by further averaging the presented results over this effect via Gaussian-Hermite quadratures.

2 Models and Problem Formulation

2.1 Fading and System Models

We consider Nakagami and Rician types of fading. Let s_D denote the instantaneous faded power of desired signal and let $\Omega_D = E_{s_D}[s_D]$ denote its short-term average. We assume that s_D is either subject to Nakagami (with Nakagami parameter m_D) or Rician (with Rician factor K_D) type of fading depending on the scenario of interest. For Nakagami fading, the probability density function (PDF) of s_D is given by

$$p_{s_D}(s_D) = \left(\frac{m_D}{\Omega_D}\right)^{m_D} \frac{s_D^{m_D-1}}{\Gamma(m_D)} \exp\left(-\frac{m_D s_D}{\Omega_D}\right), \tag{1}$$

while for Rician fading, the PDF becomes

$$p_{s_D}(s_D) = \frac{K_D+1}{\Omega_D} \exp\left(-K_D - \frac{K_D+1}{\Omega_D} s_D\right) I_0\left(2\sqrt{\frac{K_D(K_D+1)}{\Omega_D} s_D}\right), \tag{2}$$

where $I_k(\cdot)$ is the kth order modified Bessel function. Note that both $m_D = 1$ for Nakagami fading and $K_D = 0$ for Rician fading lead to Rayleigh case with the PDF given by $p_{s_D}(s_D) = \exp(-s_D/\Omega_D)/\Omega_D$.

On the other hand, the N_I active interfering signal powers $\{s_i\}_{i=1}^{N_I}$ are assumed to be IID with the same short-term average fading power Ω_I and subject to Nakagami (with Nakagami parameter m_I) or Rician (with Rician factor K_I) type of fading. The PDF of interfering signal $p_{s_i}(s_i)$ is of the same form as in (1) and (2) for the desired signals, but with different parameters. We assume throughout our study that the co-channel interfering signals add up incoherently since this leads to a more realistic assessment of the co-channel interference in cellular systems [9]. Under this assumption the total instantaneous interference power s_I is just the sum of the instantaneous powers of the N_I active interferers, i.e., $s_I = \sum_{i=1}^{N_I} s_i$.

2.2 Outage Probability

In the systems that are both interference limited and power limited, an outage is declared when either the CIR $\lambda = s_D/s_I$ falls below a predetermined protection ratio λ_{th} or the instantaneous power of desired signal s_D is less than a predetermined threshold s_{th}. Mathematically speaking, the outage probability P_{out} is thus given by

$$P_{\text{out}} = P\left[\lambda = \frac{s_D}{s_I} \le \lambda_{\text{th}} \text{ or } s_D \le s_{\text{th}}\right], \tag{3}$$

which can be written in terms of the PDF of s_D and the PDF of the total instant interference power s_I, $p_{s_I}(s_I)$, as

$$P_{\text{out}} = 1 - \int_{s_{\text{th}}}^{\infty} \left[\int_0^{s_D/\lambda_{\text{th}}} p_{s_I}(s_I)\, ds_I\right] p_{s_D}(s_D)\, ds_D. \tag{4}$$

3 Rician/IID Nakagami Scenario

3.1 Rician/IID Rayleigh Scenario

Let us assume that the received signal from desired user is Rician distributed with parameters Ω_D and K_D and that there are N_I IID Rayleigh interferers in the system with average fading power Ω_I.

3.1.1 General Case

In this case, the total instantaneous interference power s_I has a gamma PDF with the following form

$$p_{s_I}(s_I) = \frac{s_I^{N_I-1}}{\Omega_I^{N_I}(N_I-1)!}\exp\left(-\frac{s_I}{\Omega_I}\right). \tag{5}$$

Applying the finite sum representation of the gamma cumulative distribution function (CDF) and (2) in (4) gives, after switching order of integration and summation, the outage probability as

$$P_{\text{out}} = 1 - Q_1\left(\sqrt{2K_D}, \sqrt{\frac{2(K_D+1)s_{\text{th}}}{\Omega_D}}\right) + \frac{K_D+1}{\Omega_D}\sum_{n=0}^{N_I-1}\frac{1}{n!} \tag{6}$$

$$\times \int_{s_{\text{th}}}^{\infty}\left(\frac{s_D}{\lambda_{\text{th}}\Omega_I}\right)^n \exp\left[-K_D - \left(\frac{1}{\lambda_{\text{th}}\Omega_I} + \frac{1+K_D}{\Omega_D}\right)s_D\right] I_0\left(2\sqrt{\frac{K_D(K_D+1)s_D}{\Omega_D}}\right) ds_D.$$

Using an appropriate change of variable in the above integral (6), it can be shown, after some manipulations and with the help of [10], that the outage probability can be written in the following closed-form

$$P_{\text{out}} = 1 - Q_1\left(\sqrt{2K_D}, \sqrt{\frac{2(K_D+1)s_{\text{th}}}{\Omega_D}}\right) + \frac{a^2}{2K_D}\sum_{n=0}^{N_I-1}\frac{A_n}{n!}Q_{2n+1,0}(a,b), \tag{7}$$

where

$$A_n = \left(2 + \frac{2(K_D+1)\lambda_{\text{th}}\Omega_I}{\Omega_D}\right)^{-n} \exp\left[-K_D + K_D\left(1 + \frac{\Omega_D}{(K_D+1)\lambda_{\text{th}}\Omega_I}\right)^{-1}\right],$$

$$a = \sqrt{2K_D\left(1 + \frac{\Omega_D}{(K_D+1)\lambda_{\text{th}}\Omega_I}\right)^{-1}}, \quad b = \sqrt{2\left(1 + K_D + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}\right)\frac{s_{\text{th}}}{\Omega_D}},$$

and $Q_{m,n}(.,.)$ is the Nuttall Q-function defined by [10]

$$Q_{m,n}(a,b) = \int_b^{\infty} x^m \exp\left(-\frac{x^2+a^2}{2}\right) I_n(ax)\, dx. \tag{8}$$

Note that the Nuttall Q-function reduces to the generalized Marcum Q-function given by [11]

$$Q_m(a,b) = \int_b^{\infty}\left(\frac{x}{a}\right)^{m-1}\exp\left(-\frac{a^2+x^2}{2}\right) I_{m-1}(ax)x\, dx \tag{9}$$

when $m = n+1$. In deed, comparing (9) and (8), we see that $Q_{m,m-1}(a,b) = a^{m-1}Q_m(a,b)$. In particular, when $m = 1$ and $n = 0$, the Nuttall Q function reduces to the first order Marcum Q-function, i.e., $Q_{1,0}(a,b) = Q_1(a,b)$. More generally, as claimed in [10], all Nuttall functions $Q_{m,n}(a,b)$ as defined in (8) can be expressed just in terms of first order Marcum Q-function and modified Bessel functions when $m+n$ is odd. This is exactly the case for the computations of the Nuttall Q-functions involved in (7). Hence, by applying [10, Eq. (88)], it can be shown that $Q_{2n+1,0}(a,b)$ can be recursively computed for arbitrary $n \geq 1$ by using the following pair of relations

$$Q_{2n,1}(a,b) = aQ_{2n-1,0}(a,b) + 2(n-1)Q_{2n-2,1}(a,b) + b^{2n-1}\exp[-(a^2+b^2)/2]I_1(ab), \tag{10}$$

$$Q_{2n+1,0}(a,b) = aQ_{2n,1}(a,b) + 2nQ_{2n-1,0}(a,b) + b^{2n}\exp[-(a^2+b^2)/2]I_0(ab). \tag{11}$$

3.1.2 Special Cases

- Single Rayleigh Interferer

  If there is only one Rayleigh interferer, i.e., $N_I = 1$, (7) reduces to

$$P_{\text{out}} = 1 - Q_1\left(\sqrt{2K_D}, \sqrt{\frac{2(K_D+1)s_{\text{th}}}{\Omega_D}}\right) + \frac{a^2}{2K_D}\exp\left[-K_D + \frac{a^2}{2}\right] Q_1(a,b), \quad (12)$$

  in agreement with the Yao and Sheikh result [2, Eq. (4)]. For the interference limited case (i.e. $s_{\text{th}} = 0$), (12) further simplifies to $P_{\text{out}} = \frac{a^2}{2K_D}\exp(-K_D + a^2/2)$, in agreement with the Yao and Sheikh result [2, Eq. (7)].

- Two Rayleigh Interferers

  If there are two Rayleigh interferers, the outage probability is given by

$$\begin{aligned} P_{\text{out}} &= 1 - Q_1\left(\sqrt{2K_D}, \sqrt{\frac{2(K_D+1)s_{\text{th}}}{\Omega_D}}\right) + \frac{a^2}{2K_D}[(A_0 + (2+a^2)A_1)Q_1(a,b) \\ &+ A_1 \exp\left(-\frac{a^2+b^2}{2}\right)(b^2 I_0(ab) + abI_1(ab))] \end{aligned} \quad (13)$$

  where a, b and A_n for $n = 0, 1$ are the same as in (7).

3.1.3 Numerical Examples

Figure 1 shows the outage probability, as given in (7), as function of the normalized "average" CIR $\frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}$ for different Rician parameter K_D. Note that the change in Rician parameter K_D affects significantly the outage probability. However it should also be noted that the Rician parameter affects the outage probability differently for different range of CIR. For low CIR, outage probability performance is slightly improved when K_D is smaller, while for high CIR it degrades considerably when K_D decreases. Figure 2 shows the outage probability, as given in (7), for different numbers of interferers. It is observed that as N_I increases, the outage probability performance degrades for lower CIR range and it changes little for high CIR range because the minimum power requirement dominates.

3.2 Extension to Rician/IID Nakagami Case

Consider the case when the desired user is subject to Rician fading with parameters Ω_D and K_D, whereas the N_I IID interferers are subject to Nakagami fading with fading parameter m_I and average fading power Ω_I.

3.2.1 General Case

Using the Moment Generating Function (MGF) method, it can be shown that the PDF of $s_I = \sum_{i=1}^{N_I} s_i$ is

$$p_{s_I}(s_I) = \left(\frac{m_I}{\Omega_I}\right)^{m_I N_I} \frac{s_I^{m_I N_I - 1}}{\Gamma(m_I N_I)} \exp\left(-\frac{m_I s_I}{\Omega_I}\right), \quad (14)$$

where $\Gamma(\cdot)$ is the gamma function. This means that s_I has a gamma distribution with parameters $m_I N_I$ and Ω_I/m_I. Following the same procedure as for the IID Rayleigh case, the outage probability can be found in the following desired closed-form

$$P_{\text{out}} = 1 - Q\left(\sqrt{2K_D}, \sqrt{\frac{2(K_D+1)s_{\text{th}}}{\Omega_D}}\right) + \frac{a^2}{2K_D} \sum_{n=0}^{m_I N_I - 1} \frac{A_n}{n!} Q_{2n+1,0}(a,b), \qquad (15)$$

where A_n, a and b are defined in the same way as in (7) except for using $m_I \Omega_D$ in stead of Ω_D. It should be noted that (15) generalizes ([8, Eq. (14)]) (or equivalently [12, Eq. (24)]) for an interference-limited environment and gives an alternative closed-form expression for the infinite sum representation given in [8, Eq. (13)].

3.2.2 Numerical Examples

Figure 3 shows the outage probability, as given in (15), as function of normalized average CIR $\frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}$ for different Rician parameters K_D and Nakagami parameters m_I. Note that increasing m_I has little impact on outage performance, but the increase of K_D improves the performance considerably, especially for the high CNR range.

4 Nakagami/IID Rician Scenario

4.1 Rayleigh/IID Rician Scenario

Let us assume that the received power of desired signal is Rayleigh distributed with average Ω_D and that the N_I interfering signal powers are IID Rician distributed with parameter Ω_I and K_I.

4.1.1 General Case

Under the IID assumption, the PDF of the total instantaneous interference power s_I can be obtained, using the MGF method with the notion of $\Omega_I^* = \Omega_I/(K_I+1)$, as

$$p_{s_I}(s_I) = \frac{1}{\Omega_I^*}\left(\frac{s_I}{K_I N_I \Omega_I^*}\right)^{(N_I-1)/2} \exp\left(-K_I N_I - \frac{s_I}{\Omega_I^*}\right) I_{N_I-1}\left(2\sqrt{\frac{K_I N_I s_I}{\Omega_I^*}}\right). \qquad (16)$$

Substituting (16) in (4) and using the definition of the generalized (mth-order) Marcum Q-function given in (9), it can be shown that the outage probability can be written as

$$P_{\text{out}} = 1 - \exp\left(-\frac{s_{\text{th}}}{\Omega_D}\right) + \frac{1}{\Omega_D}\int_{s_{\text{th}}}^{\infty} Q_{N_I}\left(\sqrt{2N_I K_I}, \sqrt{\frac{2s_D}{\lambda_{\text{th}}\Omega_I^*}}\right) \exp\left(-\frac{s_D}{\Omega_D}\right) ds_D. \qquad (17)$$

Using an appropriate change of variable in (17), then applying the identity [11, Eq. (13)], we obtain, after some manipulations, the final desired closed-form result for the outage probability as

$$\begin{aligned} P_{\text{out}} = {} & 1 - \exp\left(-\frac{s_{\text{th}}}{\Omega_D}\right)\left[1 - Q_{N_I}\left(\sqrt{2N_I K_I}, \sqrt{\frac{2s_{\text{th}}}{\lambda_{\text{th}}\Omega_I^*}}\right)\right] - \left(1 + \frac{\lambda_{\text{th}}\Omega_I^*}{\Omega_D}\right)^{-N_I} \qquad (18) \\ & \times \exp\left[-K_I N_I \left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I^*}\right)^{-1}\right] Q_{N_I}\left[\sqrt{2N_I K_I\left(1 + \frac{\lambda_{\text{th}}\Omega_I^*}{\Omega_D}\right)^{-1}}, \sqrt{2\frac{s_{\text{th}}}{\Omega_D}\left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I^*}\right)}\right]. \end{aligned}$$

4.1.2 Special Cases

- Interference-limited Case

  For interference limited systems (i.e. $s_{\text{th}} = 0$), simplifying (18) with $Q_m(a, 0) = 1$, the outage probability reduces to

$$P_{\text{out}} = 1 - \left[1 - \left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I^*}\right)^{-1}\right]^{N_I} \exp\left[-K_I N_I \left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I^*}\right)^{-1}\right], \tag{19}$$

  which is the limiting expression one obtains (through L'Hospital's rule) from the Tjhung *et al.*'s formula [6, Eq. (13)] when setting $K_D = 0$ in it.

- Rayleigh Interferers

  When the interferer is assumed to be Rayleigh faded, we have $K_I = 0$. By applying the following property of generalized Marcum Q-function [13, Eqn. (9)] $Q_m(0, b) = \exp(-b^2/2) \sum_{n=0}^{m} (b^2/2)^n / n!$, it can be shown that (18) simplifies to

$$\begin{aligned} P_{\text{out}} &= 1 - \exp\left(-\frac{s_{\text{th}}}{\Omega_D}\right) + \exp\left[-\frac{s_{\text{th}}}{\Omega_D}\left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}\right)\right] \sum_{n=1}^{N_I - 1} \frac{1}{n!}\left(\frac{s_{\text{th}}}{\lambda_{\text{th}}\Omega_I}\right)^n \\ &- \exp\left[-\frac{s_{\text{th}}}{\Omega_D}\left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}\right)\right] \left(1 + \frac{\lambda_{\text{th}}\Omega_I}{\Omega_D}\right)^{-N_I} \sum_{n=1}^{N_I - 1} \frac{1}{n!}\left(\frac{s_{\text{th}}}{\Omega_D}\right)^n \left(1 + \frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}\right)^n . \end{aligned} \tag{20}$$

  The closed-form result (21) is equivalent to the recursive solution offered by Sowerby and Williamson in [1, Eq. (12) and Eq. (14)].

4.1.3 Numerical Examples

Figure 4 shows the outage probability, as given in (18)), as function of normalized average CIR $\frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}$ for different Rician parameters K_I. Note that, unlike the Rice/IID Rayleigh scenario, the change in K_I has little impact on the outage probability. But, it should be further noted, that similar to the Rician/IID Rayleigh case, the Rician parameter affects outage probability differently for different range of CIR. For low CIR, outage probability performance is slightly improved when K_I gets smaller, while for high CIR it surprisingly degrades when K_I decreases. Figure 5 shows the outage probability, as given in (18), for different numbers of interferers N_I. It is observed that as N_I increases, the outage probability performance degrades significantly for lower CIR range and it changes little for high CIR range since the minimum power requirement dominates, as it was the case for the Rician/IID Rayleigh case, but in this unfavorable scenario the outage probabilities are much higher then in the previous favorable scenario.

4.2 Extension to Nakagami/IID Rician Scenario

Consider the case when the desired user is subject to Nakagami fading with parameters Ω_D and m_D, where the N_I IID interferers are subject to Rician fading with fading parameter K_I and average fading power Ω_I.

4.2.1 General Case

Using Nakagami PDF instead of Rayleigh PDF in (17), the outage probability is

$$\begin{aligned} P_{\text{out}} &= 1-\sum_{i=0}^{m_D-1}\frac{1}{i!}\left(\frac{m_D s_{\text{th}}}{\Omega_D}\right)^i \exp\left(-\frac{m_D s_{\text{th}}}{\Omega_D}\right) \\ &+ \int_{s_{\text{th}}}^{\infty} Q_{N_I}\left(\sqrt{2N_I K_I}, \sqrt{\frac{2(K_I+1)s_D}{\lambda_{\text{th}}\Omega_I}}\right)\left(\frac{m_D}{\Omega_D}\right)^{m_D}\frac{s_D^{m_D-1}}{(m_D-1)!}\exp\left(-\frac{m_D s_D}{\Omega_D}\right) ds_D. \end{aligned} \tag{21}$$

With an appropriate change of variable and application the following definitions

$$a=\sqrt{2N_I K_I},\ b=\sqrt{\frac{(K_I+1)\Omega_D}{m_D\lambda_{\text{th}}\Omega_I}},\ c=\sqrt{\frac{2m_D s_{\text{th}}}{\Omega_D}}, \tag{22}$$

and (8), the integral in (21) denoted by R_{m_D} can be shown through integrating by part to satisfy the following recursive relationship

$$\begin{aligned} R_{m_D} &= R_{m_D-1}+\frac{(c^2/2)^{m_D-1}}{(m_D-1)!}\exp\left(-\frac{c^2}{2}\right)Q_{N_I}(a,bc)-\frac{[2(b^2+1)]^{1-m_D}}{(m_D-1)!}\left(\frac{b}{\sqrt{b^2+1}}\right)^{N_I+1} \\ &\times \left(\frac{1}{a}\right)^{N_I-1}\exp\left(-\frac{a^2}{2(b^2+1)}\right)Q_{2(m_D-1)+N_I,N_I-1}\left(\frac{ab}{\sqrt{b^2+1}},c\sqrt{b^2+1}\right). \end{aligned} \tag{23}$$

Therefore, after some manipulation, the outage probability can be written in the following closed-form

$$\begin{aligned} P_{\text{out}} &= 1-\sum_{i=0}^{m_D-1}\frac{(c^2/2)^i}{i!}\exp\left(-\frac{c^2}{2}\right)[1-Q_{N_I}(a,bc)]+\left(\frac{b}{\sqrt{b^2+1}}\right)^{N_I+1}\left(\frac{1}{a}\right)^{N_I-1} \\ &\times \exp\left(-\frac{a^2}{2(b^2+1)}\right)\sum_{i=0}^{m_D-1}\frac{[2(b^2+1)]^{-i}}{i!}Q_{2i+N_I,N_I-1}\left(\frac{ab}{\sqrt{b^2+1}},c\sqrt{b^2+1}\right), \end{aligned} \tag{24}$$

where a, b and c are given in (22). Note that (24) generalizes ([7, Eq. (13)]) (or equivalently [12, Eq. (26)]) which are only valid for an interference-limited environment. It should be further noted that since all the Nuttall Q-functions $Q_{m,n}(\cdot,\cdot)$ in (24) satisfy $m=n+2i+1$, $i=0,1,2,\cdots,m_D-1$, they can solely be expressed in terms of generalized Marcum Q-functions and modified Bessel functions by repeatedly applying the following formula [10, Eq. (87)]

$$Q_{m,n}(a,b)=aQ_{m-1,n+1}(a,b)+(m+n-1)Q_{m-2,n}(a,b)+b^{m-1}\exp\left(-\frac{a^2+b^2}{2}\right)I_{n-1}(ab). \tag{25}$$

It should be observed that the general formula given in (24) is not applicable when $a=0$, i.e. either $K_I=0$ or $N_I=0$. The case of $K_I=0$ corresponds to the Nakagami/IID Rayleigh combination, for which the outage probability can be easily obtained as a special case of the Yao and Sheikh result [3, Eq. (18)]. The case of $N_I=0$ leads to the noise limited only case.

4.2.2 Numerical Examples

Figure 6 shows the outage probability, as given in (24), as function of normalized average CIR $\frac{\Omega_D}{\lambda_{\text{th}}\Omega_I}$ for different Rician parameters K_I and Nakagami parameters m_D. Note that like in the Rician/IID Nakagami case, increasing K_I has little impact on outage performance, but the increase of m_D improves the performance considerably, especially for the high CNR range.

5 Conclusion

Two special co-channel interference fading combinations have been investigated in this paper. More specifically, with the consideration of both minimum CIR and CNR constraints, closed-form expressions for the outage probability of both the Nakagami/IID Rician and Rician/IID Nakagami combinations when the Nakagami fading parameter is restricted to integer values have been derived. The final closed-form expressions are expressed in terms of the Marcum Q-function and modified Bessel functions. When applicable, the results obtained as special cases of the offered closed-form expressions have been compared to those previously reported in the literature. In addition, as an illustration/validation of the mathematical formalism, some numerical examples have been provided and discussed for both fading combinations under consideration.

References

[1] K. W. Sowerby and A. G. Williamson, "Outage probability calculations for multiple cochannel interferers in cellular mobile radio systems," *IEE Proc. (Pt. F)*, vol. 135, pp. 208–215, June 1988.

[2] Y. D. Yao and A. U. H. Sheikh, "Outage probability analysis for microcell mobile radio systems with co-channel interferers in Rician/Rayleigh fading environment," *IEE Electron. Lett.*, vol. 26, pp. 864–866, June 1990.

[3] Y. -D. Yao and A. U. H. Sheikh, "Investigation into cochannel interference in microcellular mobile radio systems," *IEEE Trans. Veh. Technol.*, vol. VT-41, pp. 114–123, May 1992.

[4] J. R. Haug and D. R. Ucci, "Outage probability for microcellular radio systems in a Rayleigh/Rician fading environment," in *Proc. IEEE Int. Conf. on Commun. (ICC'92)*, pp. 312.4.1–312.4.5, June 1992.

[5] T. T. Tjhung, C. C. Chai, and X. Dong, "Outage probability for a Rician signal in L Rician interferers," *IEE Electron. Lett.*, vol. 31, pp. 532–533, March 1995.

[6] T. T. Tjhung, C. C. Chai, and X. Dong, "Outage probability for lognormal-shadowed Rician channels," *IEEE Trans. Veh. Technol.*, vol. VT-46, pp. 400–407, May 1997.

[7] L. C. Wang and C. T. Lea, "Co-channel interference analysis of shadowed Rician channels," *IEEE Commun. Letters*, vol. 2, pp. 67–69, March 1998.

[8] J. -C. Lin, W. -C. Kao, Y. T. Su, and T. -H. Lee, "Outage and coverage considerations for microcellular mobile radio systems in a shadowed-Rician/shadowed-Nakagami environment," *IEEE Trans. Veh. Technol.*, vol. VT-48, pp. 66–75, January 1999.

[9] R. Prasad and A. Kegel, "Improved assessement of interference limits in cellular radio performance," *IEEE Trans. Veh. Technol.*, vol. VT-40, pp. 412–419, May 1991.

[10] A. H. Nuttall, "Some integrals involving the Q function." Research Report of Naval Underwater Systems Center, 1972.

[11] A. H. Nuttall, "Some integrals involving the Q_m function," *IEEE Trans. Info. Theory*, pp. 95–96, January 1975.

[12] M. K. Simon and M. -S. Alouini, "On the difference of two chi-square variate with application to outage probability computation," in *Proc. of the 38th Allerton Conference on Communication, Control, and Computing (Allerton'2000), Monticello, Illinois*, October 2000.

[13] G. M. Dillard, "Recursive computation of the generalized Q-function," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-9, pp. 614–615, July 1973.

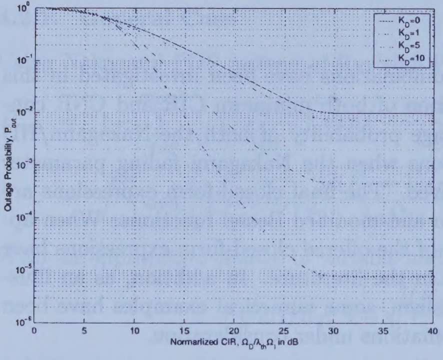


Figure 1: Outage Probability P_{out} as function of the normalized CIR $\Omega_D/\lambda_{\text{th}}\Omega_I$ for Rician/IID Rayleigh scenario with $N_I = 6$ and $\Omega_D/s_{\text{th}} = 20$ dB.

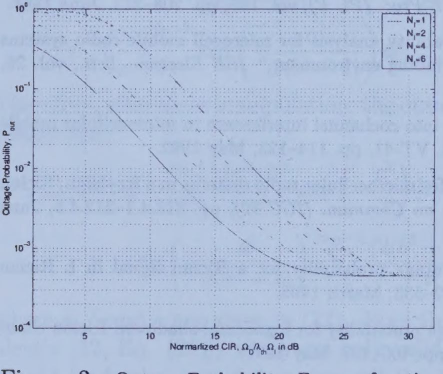


Figure 2: Outage Probability P_{out} as function of the normalized CIR $\Omega_D/\lambda_{\text{th}}\Omega_I$ for Rician/IID Rayleigh scenario with $K_D = 5$ and $\Omega_D/s_{\text{th}} = 20$ dB.

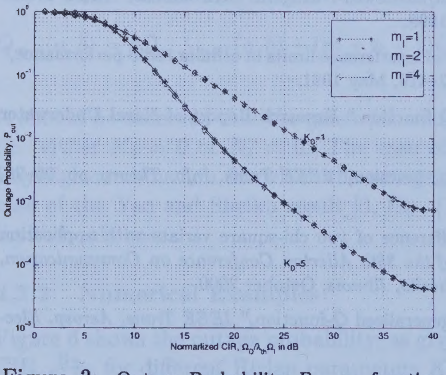


Figure 3: Outage Probability P_{out} as function of the normalized CIR $\Omega_D/\lambda_{\text{th}}\Omega_I$ for Rician/IID Nakagami scenario with $N_I = 6$ and $\Omega_D/s_{\text{th}} = 30$ dB.

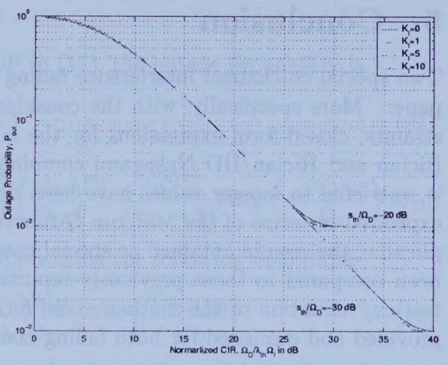


Figure 4: Outage Probability P_{out} as function of the normalized CIR $\Omega_D/\lambda_{\text{th}}\Omega_I$ for Rayleigh/IID Rician scenario with $N_I = 6$ and $\Omega_D/s_{\text{th}} = 20$ and 30 dB.

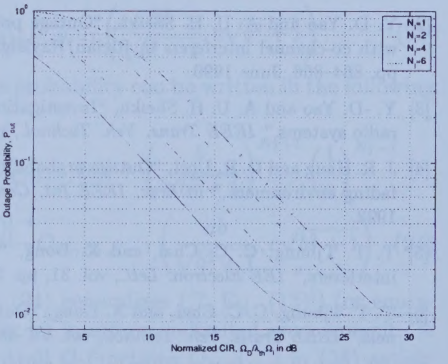


Figure 5: Outage Probability P_{out} as function of the normalized CIR $\Omega_D/\lambda_{\text{th}}\Omega_I$ for Rayleigh/IID Rician scenario with $K_I = 5$ and $\Omega_D/s_{\text{th}} = 20$ dB.

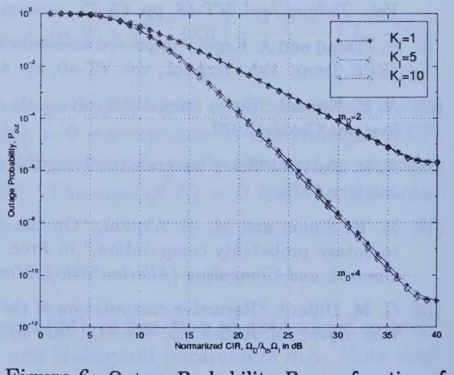


Figure 6: Outage Probability P_{out} as function of the normalized CIR $\Omega_D/\lambda_{\text{th}}\Omega_I$ for Nakagami/IID Rician scenario with $N_I = 6$ and $\Omega_D/s_{\text{th}} = 30$ dB.

A Novel Hybrid Single-Carrier/Multicarrier System for Highly Bandwidth Efficient Wireless Communication*

Ganesh Ananthaswamy and Dennis L. Goeckel
Dept. of Electrical and Computer Engineering
University of Massachusetts
Amherst, MA 01003-5110
{ganantha,goeckel}@ecs.umass.edu

Abstract

An orthogonal frequency division multiplexing (OFDM) system operating over a wireless communication channel exhibits two weaknesses relative to its single-carrier counterparts: (1) the diversity achieved by the OFDM system can be less than a single-carrier system employing the same error control code in a signaling environment rich in diversity, and (2) the baseband transmitted signal can exhibit significant amplitude fluctuation over time, thereby precluding efficient transmit amplifier operation. In [1] and [2], the desire to reduce both of these impairments motivated a novel coded modulation construction for OFDM systems. In effect, this construction was incorporating some elements of a single-carrier system into the OFDM system, thus forming a single-carrier/multicarrier hybrid. In this paper, we explore the utility of this construction as an alternative to the most commonly considered single-carrier/multicarrier hybrid, which breaks the total bandwidth up into a number of single-carrier systems. To facilitate this comparison, a novel reduced complexity decoding algorithm is developed for the proposed hybrid. Numerical results are presented that demonstrate the superiority of the proposed hybrid system over the standard hybrid system.

1 Introduction

The scarcity of usable frequency bands has motivated the development of various signalling strategies for highly bandwidth-efficient communication. A candidate system can generally be categorized as either a single-carrier (SC) system or a multi-carrier (MC) system. SC systems make use of the entire available bandwidth for each channel symbol, correcting for any distortion inherent to a frequency-selective channel by employing some form of equalization. MC systems deal with frequency selectivity by breaking up the available spectrum into several smaller frequency non-selective channels and then transmitting narrowband signals on each of these subchannels, thereby obviating the need for an equalizer. This paper emphasizes the hybrid SC-MC nature of the system presented in [1],[2] (henceforth called System A) and compares its performance with the

---

*This research is supported in part by the National Science Foundation under grant NCR-9714597 and CAREER Award CCR-9875482.

most commonly considered hybrid system (System B), which will be described shortly. System A is obtained by modifying a fast Fourier transform (FFT) implementation of the conventional OFDM system, whereas System B is obtained by dividing up the entire available spectrum into a number of contiguous subchannels, each of which is treated as a SC system. A suitable channel code is employed in each of the systems and their performance is compared at the same overall channel diversity for various numbers of code states. By construction, a channel symbol in system A employs subcarriers spread broadly over the system bandwidth, whereas the bandwidth occupied by a channel symbol in system B is contiguous. Herein lies the advantage of System A; for a given receiver complexity and peak-to-mean envelope power ratio (PMEPR), each channel symbol is able to avail itself of a greater amount of frequency diversity because it is spread out across the entire available bandwidth, whereas a channel symbol in System B, with its frequency band confined to a contiguous part of the spectrum, is unable to capture this frequency diversity. Furthermore, if the channel delay spread is not known precisely at the receiver *a priori*, the equalizer of the MMSE-DFE in System B could have a poor performance versus complexity tradeoff due to a mis-estimate of the number of taps to employ in the feedforward and the feedback filters.

The complexity of the bit metric generator in System A, as presented in [1],[2] increases as $2^{M \cdot L}$, where L is the number of bits per symbol in the constituent signal set and M is the number of complex dimensions of the multidimensional signal, implying that the decoder complexity increases exponentially with M and L. Hence, if high diversity (i.e. large M) and/or high spectral efficiency (i.e. large L) is desired, very high decoding complexity would have to be tolerated and hence there is a need for low-complexity decoding algorithms. A low-complexity bit metric generation algorithm for System A is developed in this paper. The basic idea is to consider only those signals that lie within a certain predetermined radius of the received signal. However, the task of enumerating the signals that lie within this predetermined radius is non-trivial when the underlying lattice, to which the transmitted signal belongs, is non-uniform. For instance, it is easy to determine the point in an integer lattice (say $\mathcal{Z}^{2M}$) closest to a received signal, but it is difficult when the integer lattice has been transformed by a unitary matrix, particularly when the transmitted signal is corrupted by additive noise. Prior work in this area appeared in [4], and an alternate derivation of the algorithm presented in [4] has been proposed in [5]. The algorithm employs a Cholesky decomposition of the generating matrix of the lattice to obtain a recursive form that essentially enumerates all of the lattice points within a certain radius of the received signal vector. While the algorithm of [4] works well for signals corrupted by additive white Gaussian noise (AWGN), it does not perform as well in fading channels, since the number of signal points within a fixed predetermined radius of the received signal varies as the fade level changes. For instance, if the fading amplitude is very low, there are many more points within the fixed radius as compared to the case when the fading amplitude is high. This means that there are many signal points contributing to a bit metric when the value of the fading is small but perhaps not many signals contributing to the bit metric when the value of the fading is large. This results in suboptimal performance for a given complexity, which is taken here to be proportional to the average number of signal points. Also, the variation in the number of signal points required to compute a bit metric directly translates into a variation in the number of operations required to compute a bit metric, which can cause difficulties in highly specialized receiver circuit designs that operate in a parallel rather than in a sequential manner. In order to avoid this problem, the radius of the hyper-

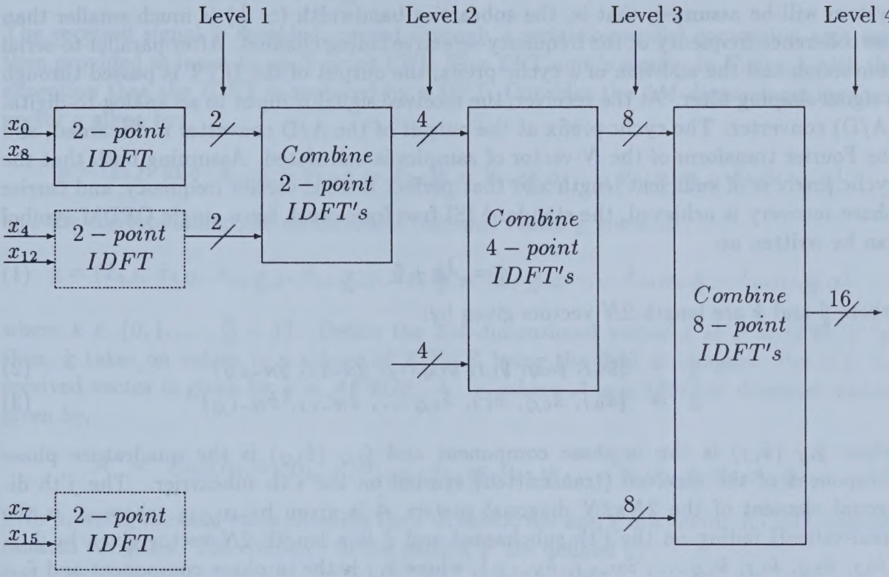


Figure 1: Block diagram of System A showing the construction of a 16-point IFFT as a combination of several lower order IFFT's. The index on x, the input to the IFFT, is the subcarrier on which the corresponding symbol is eventually placed. The construction of the FFT in the receiver is identical except that the IDFT's above are replaced with DFT's.

sphere needs to be adapted based on the fade value of the received signal. This point has also been made in [5], but no adaptation method has been outlined. A modified algorithm based on a complexity analysis presented in [4] is used to adapt the radius of the hypersphere in an attempt at maintaining the number of signal points within the hypersphere a constant.

The organization of this paper is as follows: The construction of the proposed hybrid is presented in Section 2.1, and the receiver processing that would be required for the optimal bit metric generation based on the received signal is presented in Section 2.2. Section 2.3 analyzes the decoder complexity, while Section 2.4 presents a modified reduced complexity algorithm to generate bit metrics in System A. Finally, the performance of the proposed system is compared with that of the standard hybrid in Section 3.

2 System Description

2.1 Proposed Hybrid

One popular approach to the implementation of a conventional OFDM system is by means of an IFFT [6]. The approach consists of gathering the symbols to be transmitted into blocks of length equal to the number of subcarriers (N) in the OFDM system and performing an IFFT on each of these blocks. The key assumption that results in effectively a frequency-nonselective fading channel for the coded symbols in an OFDM system will be assumed; that is, the subcarrier bandwidth ($\simeq \frac{1}{T}$) is much smaller than the coherence frequency of the frequency-selective fading channel. After parallel-to-serial conversion and the addition of a cyclic prefix, the output of the IFFT is passed through a signal shaping filter. At the receiver, the received signal is input to an analog-to-digital (A/D) converter. The cyclic prefix at the output of the A/D converter is discarded, and the Fourier transform of the N-vector of samples is calculated. Assuming both that the cyclic prefix is of sufficient length and that perfect timing, carrier frequency, and carrier phase recovery is achieved, the standard ISI-free framework for a single OFDM symbol can be written as:

$$\underline{\hat{y}} = \hat{\mathcal{A}}\underline{\hat{x}} + \underline{\hat{\nu}} \tag{1}$$

where $\underline{\hat{y}}$ and $\underline{\hat{x}}$ are length-$2N$ vectors given by:

$$\begin{aligned} \underline{\hat{y}} &= \{\hat{y}_{0,I},\ \hat{y}_{0,Q},\ \hat{y}_{1,I},\ \hat{y}_{1,Q}\cdots,\ \hat{y}_{N-1,I},\ \hat{y}_{N-1,Q}\} && (2)\\ \underline{\hat{x}} &= \{\hat{x}_{0,I},\ \hat{x}_{0,Q},\ \hat{x}_{1,I},\ \hat{x}_{1,Q}\cdots,\ \hat{x}_{N-1,I},\ \hat{x}_{N-1,Q}\} && (3)\end{aligned}$$

where $\hat{y}_{i,I}$ ($\hat{x}_{i,I}$) is the in-phase component and $\hat{y}_{i,Q}$ ($\hat{x}_{i,Q}$) is the quadrature phase component of the received (transmitted) symbol on the i'th subcarrier. The j'th diagonal element of the $2N$x$2N$ diagonal matrix $\hat{\mathcal{A}}$ is given by $\alpha_{\lfloor j/2 \rfloor}$, where α_i is the (real-valued) fading on the i'th subchannel and $\underline{\hat{\nu}}$ is a length $2N$-vector given by $\underline{\hat{\nu}} = \{\hat{\nu}_{0,I},\ \hat{\nu}_{0,Q},\ \hat{\nu}_{1,I},\ \hat{\nu}_{1,Q}\dots,\ \hat{\nu}_{N-1,I},\ \hat{\nu}_{N-1,Q}\}$, where $\hat{\nu}_{i,I}$ is the in-phase component and $\hat{\nu}_{i,Q}$ is the quadrature phase component of the noise on the i'th subchannel.

The block diagram in Figure 1 shows the construction of a 16-point IFFT by the decimation-in-input construction, which consists of a cascade of several lower order IFFT's [7]. In the first level, two 2-point IFFT's are combined to form each 4-point IFFT. In the second level, two 4-point IFFT's are combined to form each 8-point IFFT, etc. Figure 1 facilitates an interesting view leading to a hybrid SC-MC system. The conventional OFDM system has a 2^L-ary QAM symbol on each input of the IFFT while the conventional SC system has a 2^L-ary QAM symbol on each output of the IFFT (in practice, of course, no IFFT or cyclic prefix is required in the case of the SC system). The hybrid System A is formed by designing the transmitted signal set at the input to the N-point IFFT in such a way that each symbol at a predetermined level of the IFFT is drawn from a rotated 2^L-ary QAM constellation. The predetermined level is the output of the M-point IFFT which forms a building block of the N-point IFFT. For instance: if $M = 4$, the input to the eight 2-point IFFT's in Figure 1 is specified such that each output at level 2 is drawn from a rotated 2^L-ary QAM constellation symbol: $(x_k, x_{k+8}, x_{k+4}, x_{k+12}) \in \{\mathcal{F}_4(\overline{\Phi}\underline{a}) : \underline{a} \in \mathcal{J}^4\}$ where $\mathcal{J}$ is the constituent QAM constellation, $k \in \{0, 1, 2, 3\}$, $\overline{\Phi}$ is an MxM diagonal matrix whose k'th element is given by $e^{j\phi_k}$, where $\{\phi_k,\ k = 0, 1, 2, 3\}$ is a set of appropriate rotations [2] determined to minimize the bit error rate (BER) of the system, and $\mathcal{F}_4(.)$ is the 4-point DFT of its argument. This construction has two consequences: (a) it leads to a lower system PMEPR compared to the conventional OFDM system, and (b) because of the method in which the rotations are chosen and because the inputs to an M-point IDFT correspond to subcarriers equally spread across the system bandwidth, (e.g. for the specific case of $M = 4$ and $N = 16$, $(\hat{x}_k, \hat{x}_{k+8}, \hat{x}_{k+4}, \hat{x}_{k+12})$ are spread evenly across the system bandwidth), the proposed hybrid obtains a higher diversity for each channel symbol compared to an OFDM system, but at the expense of a higher bit metric complexity as detailed below.

2.2 Receiver Processing

The received signal is sampled, passed through a serial-to-parallel conversion unit and then provided as input to an N-point FFT. This FFT unit is similar to Figure 1 with the exception that the IDFT is replaced by a DFT. Consider the $2M$-dimensional received vector $\underline{y}$ given by:

$$\underline{y} = (\hat{y}_{k,I},\ \hat{y}_{k,Q},\ \hat{y}_{k+\frac{N}{M},I},\ \hat{y}_{k+\frac{N}{M},Q},\ \hat{y}_{k+\frac{2N}{M},I},\ \hat{y}_{k+\frac{2N}{M},Q}, \ldots,\ \hat{y}_{k+N-\frac{N}{M},I},\ \hat{y}_{k+N-\frac{N}{M},Q})$$

and the corresponding $2M$-dimensional transmit vector $\underline{x}$ given by

$$\underline{x} = (\hat{x}_{k,I},\ \hat{x}_{k,Q},\ \hat{x}_{k+\frac{N}{M},I},\ \hat{x}_{k+\frac{N}{M},Q},\ \hat{x}_{k+\frac{2N}{M},I},\ \hat{x}_{k+\frac{2N}{M},Q}, \ldots,\ \hat{x}_{k+N-\frac{N}{M},I},\ \hat{x}_{k+N-\frac{N}{M},Q})$$

where $k \in \{0, 1, \ldots, \frac{N}{M} - 1\}$. Define the $2M$-dimensional vector $\tilde{\underline{x}}$ as $\tilde{\underline{x}} = (F\Phi\tilde{G})^{-1}\underline{x}$; then, $\tilde{\underline{x}}$ takes on values in a subset of $\mathcal{Z}^{2M}$, $\mathcal{Z}$ being the field of integers. Per (1), the received vector is given by $\underline{y} = \mathcal{A}F\Phi\tilde{G}\tilde{\underline{x}} \ + \ \underline{\nu}$, where $\mathcal{A}$ is a $2M$x$2M$ diagonal matrix given by,

$$\mathcal{A} \ = \ \text{diag}(\alpha_k, \alpha_k, \alpha_{k+\frac{N}{M}}, \alpha_{k+\frac{N}{M}}, \alpha_{k+\frac{2N}{M}}, \alpha_{k+\frac{2N}{M}}, \ldots, \alpha_{k+N-\frac{N}{M}}, \alpha_{k+N-\frac{N}{M}}) \tag{4}$$

with α_i being the fade value effecting the i'th subcarrier and $\underline{\nu}$ is a vector of i.i.d Gaussian random variables. The elements of the matrix F are defined by

$$F_{n,k} \triangleq \frac{1}{\sqrt{N}}\mathcal{R}\left(e^{-j\frac{2\pi nk}{N}}\right) \tag{5}$$

where, by definition,

$$\mathcal{R}(x) \ \triangleq \ \begin{pmatrix} Re(x) & -Im(x) \\ Im(x) & Re(x) \end{pmatrix} \tag{6}$$

and $F_{n,k}$ is the $(n,k)^{th}$ element of the block Fourier transform matrix F. Define the block diagonal rotation matrix as $\Phi \triangleq \text{diag}(\mathcal{R}(e^{j\phi_0}), \mathcal{R}(e^{j\phi_1}), \ldots, \mathcal{R}(e^{j\phi_{M-1}}))$ and $\tilde{G}$ as the 'generating' matrix [8] of the lattice. Since this work focusses on 2^L-ary QAM constellations only (with signal points lying on a lattice whose coordinates are all odd), the generating matrix takes on the form of the identity matrix and the equation for the received signal is modified as:

$$\underline{y} = 2\mathcal{A}F\Phi\tilde{\underline{x}} \ - \ \mathcal{A}F\Phi[1, 1, \ldots, 1]^T \ + \underline{\nu} \tag{7}$$

This would ensure that, even as the components of the vector $\tilde{\underline{x}}$ take on all integer values (within some bounds), the transmitted vector corresponds to the desired multidimensional QAM constellation.

2.3 The decoding complexity

The coded modulation scheme considered in this paper is bit-interleaved coded modulation (BICM) [3], which is chosen for the robust performance that it exhibits in channels ranging from AWGN channels to Rayleigh fading channels. In a BICM system, the bits to be transmitted are first convolutionally encoded, interleaved at the bit level and then the interleaved sequence is split into blocks of length $M \cdot L$. The block of bits are then used to pick an M complex-dimensional symbol for transmission. Let $l^i(\underline{x})$ denote the i^{th} bit of the label of the $2M$-dimensional vector $\underline{x}$. If the $2M$-dimensional vector $\underline{y}$ is the channel output resulting from the transmission of $\underline{x}$, and $\mathcal{X}_b^i$ denotes the set of all symbols whose label has b as the i^{th} bit, then, the bit metric $\lambda^i(\underline{y}, b)$, is given by:

$$\lambda^i(\underline{y}, b) = \log\left(\frac{1}{|\mathcal{X}_b^i|}\sum_{\underline{x}\in\mathcal{X}_b^i} p_{\mathcal{A}}(\underline{y}/\underline{x})\right) \tag{8}$$

where $p_{\mathcal{A}}(\cdot)$ is the probability density function of the argument assuming that perfect channel state information $\mathcal{A}$ (i.e. perfect knowledge of the fade values α_i effecting all the subcarriers) is available at the receiver. Irrespective of whether the M subcarriers, which are equally spaced across the bandwidth, are independently faded or not, the conditional probability density function $p_{\mathcal{A}}(\cdot)$ is given by:

$$p_{\mathcal{A}}(\underline{y}/\underline{x}) = \frac{1}{\sqrt{2^M\pi\sigma_\nu^{2M}}}\, e^{-\frac{\|\underline{y}-\mathcal{A}\underline{x}\|^2}{2\sigma_\nu^2}} \tag{9}$$

where σ_ν^2 is the variance of a component of $\underline{\nu}$. It is clear from the above equation that the bit metric is a function of all of the symbols in $\mathcal{X}_b^i$. Since the number of such distances to be calculated increases exponentially with the number of information bits in a constituent signal set and the dimensionality of the signal constellation, approximate bit metrics are generally employed. This consists of determining the set of points $\mathcal{X}_\gamma(\underline{y})$ that are within a predetermined radius (γ) of the received signal $\underline{y}$ and choosing only those points that belong to $\mathcal{X}_\gamma(\underline{y}) \cap \mathcal{X}_b^i$ in the summation in (8); in other words, the approximate bit metric

$$\hat{\lambda}^i(\underline{y}, b) = \log\left(\frac{1}{|\mathcal{X}_b^i|}\sum_{\underline{x}\in\mathcal{X}_b^i\cap\mathcal{X}_\gamma(\underline{y})} p_{\mathcal{A}}(\underline{y}/\underline{x})\right) \tag{10}$$

is defined.

While the idea behind this approximate algorithm is easily stated, the problem lies in the ability to determine the points that lie within a predetermined radius on a non-uniform lattice, as is the case in the current context. Moreover, as mentioned earlier, the number of such signal points changes with the fade value if the radius of the hypersphere within which the signal points are enumerated is held fixed. This necessitates an adaptation criterion for the radius of the said hypersphere. The problem of determining signal points within a hypersphere centered at an arbitrary point is identical to the problem of determining vectors of short length in a lattice (see [4] and [5]), as will be seen in the next section. A modified version of the the algorithm first proposed in [4] and later in [5] is discussed in the next section.

2.4 The reduced complexity algorithm

As per (7) and (9), the bit metric for a candidate transmitted symbol $\tilde{\underline{x}}$ is evaluated as follows:

$$\begin{aligned} \|\underline{y} - 2\mathcal{A}F\Phi\tilde{\underline{x}} + \mathcal{A}F\Phi[1,1,\ldots,1]^T\| &= \|(\underline{y} + \mathcal{A}F\Phi[1,1,\ldots,1]^T) - 2\mathcal{A}F\Phi\tilde{\underline{x}}\| & (11)\\ &= \|\tilde{\underline{y}} - 2\mathcal{A}F\Phi\tilde{\underline{x}}\| & (12) \end{aligned}$$

where $\tilde{\underline{y}} \triangleq \underline{y} + \mathcal{A}F\Phi[1,1,\ldots,1]^T$ is the "corrected" received vector. Define the $2M$x$2M$ matrix $\tilde{M}$ as $\tilde{M} \triangleq 2\mathcal{A}F\Phi$. Then, the equivalent transmitted symbol is given by $\tilde{M}\tilde{\underline{x}}$

where $\tilde{x}$ is a $2M$-dimensional vector of integers (i.e. $\tilde{x} \in \mathcal{Z}^{2M}$) whose i'th component is denoted as $\tilde{x}_i$. The algorithm presented in [5] may now be employed. A complexity analysis of the above algorithm has been carried out in [4] where it is shown that an upperbound κ on the number of constellation points enumerated by the algorithm is:

$$\kappa = \left(2\left\lfloor\sqrt{\gamma/q_{\min}}\right\rfloor + 1\right) \cdot \binom{\lfloor 4\gamma/q_{\min}\rfloor + 2M - 1}{\lfloor 4\gamma/q_{\min}\rfloor} \tag{13}$$

where $q_{\min}$ is a lower bound to the eigenvalues of the Gram matrix $\tilde{M}^T\tilde{M}$ and γ is the radius of the hypersphere. It is clear from the above expression that the number of signal points that would be enumerated by the algorithm depends on the product of the radius γ of the hypersphere and $1/q_{\min}$. The only term in the above expression that depends on the channel is $q_{\min}$. Hence, the proposed modification to the algorithm controls the number of signal points enumerated by the algorithm by multiplying the predetermined radius γ by q_{min}. It is further recognized that the AWGN may result in the received signal point spreading further out as the SNR decreases, and hence the radius is corrected by adding a factor equal to the total noise variance i.e. $\gamma = Cq_{min} + 2M\sigma_\nu^2$, where C is an appropriate constant.

3 Numerical Results

3.1 Overview of System B

Consider a SC system supporting a data rate of N/T symbols per unit time - the same as that of a conventional N-carrier OFDM system of symbol period T. The baseband representation of the symbol spaced channel response of the M-tap channel over which the SC system operates is given by:

$$h(t,\tau) = \sum_{k=0}^{M-1} h_k(t)\delta\left(\tau - \frac{kT}{N}\right) \tag{14}$$

where $h(t,\tau)$ is the response at time t due to an impulse applied at time '$t-\tau$', and $h_k(t)$ are statistically independent complex Gaussian random variables. Assuming Nyquist (or "sinc") pulse shaping at the transmitter, the effective instantaneous frequency spectrum of the channel response is given by:

$$H(t,f) = \sum_{k=0}^{M-1} h_k(t)e^{-j\frac{2\pi fkT}{N}} \qquad -\frac{N}{2T} \le f \le \frac{N}{2T} \tag{15}$$

System B is formed by dividing up the entire available spectrum into $Q = N/M$ subchannels and a SC system now operates independently on each subchannel. The data at the transmitter after channel coding is interleaved (in order to obtain time diversity) and then mapped onto a signal set. The data is shaped by Nyquist pulses, transmitted, and sampled at the symbol rate at the receiver. The baseband representation, $h_l(t,\tau)$, of the l'th subchannel can be shown to be given by:

$$\begin{aligned} h_l(t,\tau) &= \int_{-\frac{N}{2TQ}}^{\frac{N}{2TQ}} \sum_{k=0}^{M-1} h_k(t)e^{-j2\pi\left(f-\frac{N}{2T}+\frac{(2l-1)N}{2TQ}\right)\frac{kT}{N}} e^{j2\pi f\tau}df && (16) \\ &= \sum_{k=0}^{M-1} h_k'(t)\left[\left(\left(\frac{N}{TQ}\right)\frac{\sin\left(\frac{\pi N}{TQ}\left(\frac{kT}{N}-\tau\right)\right)}{\left(\frac{\pi N}{TQ}\left(\frac{kT}{N}-\tau\right)\right)}\right)\right] && (17) \end{aligned}$$

SNR (dB)	$QPSK^4$	$QPSK^8$	$16-QAM^4$
2	80.371003	8562.762641	-
3	65.238691	5278.744049	-
4	52.130956	3173.166282	2219.886770
5	41.143425	1769.430750	1471.695917
6	32.248650	949.813864	970.145325
7	25.273137	500.550351	637.072030
8	19.859460	258.533919	421.494675
9	-	-	281.958050
10	-	-	193.997395

Figure 2: A table illustrating the average number of signal points that are taken into account when evaluating the bit metric of equation (10) at different SNR's and for different transmit constellations.

where $h_k^{'}(t)$ has the same statistics as $h_k(t)$. The symbol rate sampled channel at delay 'm' is obtained by sampling the above response at $\tau = \frac{mQT}{N}$, where 'm' is an integer. In the simulations, all the $h_k(t)$ are assumed to be independent zero mean unit variance complex Gaussian random variables. This would ensure that the M-tap channel model would have a frequency diversity of M. The received waveform is sampled at the symbol rate and passed through a whitened matched filter. The output of the matched filter is passed through a deinterleaver, equalized with an MMSE-DFE and decoded. The simulations assume that the receiver has perfect knowledge of the channel and that there are no adaptation errors.

3.2 Performance comparison of System A with System B

Simulation results presenting the number of constellation points enumerated by the reduced complexity algorithm for the case of (a) four-dimensional QPSK, (b) an eight-dimensional QPSK and (c) a four-dimensional 16-QAM transmit constellation are shown in Figure 2. If the optimal metric were to be used, the number of signal points to be summed up in (8) would be 256 for (a) and 65536 for each of (b) and (c). However, using the reduced complexity algorithm, the number of signal points evaluated at different signal to noise ratio (SNR) is as shown in Figure 2. The huge savings in complexity is apparent. The number of signal points enumerated is seen to decrease with an increase in SNR, as is to be expected by the way in which the radius γ of the hypersphere is adapted.

The required average transmit SNR for System A (obtained using the reduced complexity algorithm above) and System B is shown in Figure 3. It is seen that at the same overall channel diversity, the performance of System A is superior to the performance of System B. Perfect knowledge of the channel as well as zero equalizer adaptation error have been assumed. The plots correspond to a bit error rate of 10^{-5}. The performance of System A as compared to System B is expected to get better at lower bit error rates since the effect of the increased diversity available to System A is prominent at lower bit error rates. The advantage in employing high dimensional constellations (provided of course that an equal amount of frequency diversity is available) is seen from both Figure 3 and

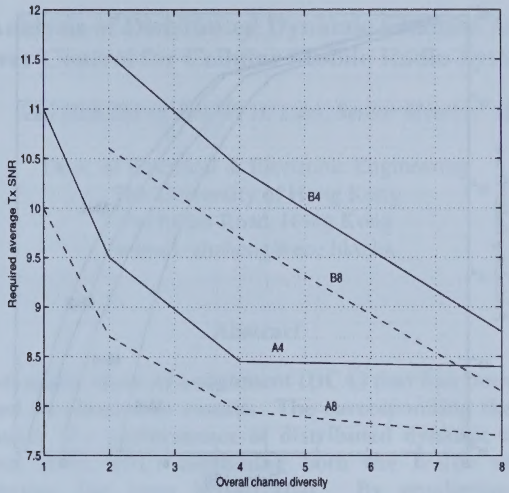


Figure 3: The average transmit SNR required by System A and System B to achieve a bit error rate of 10^{-5} at various overall channel diversities and different number of code states. The label Ax indicates System A using an x state channel code while Bx is its counterpart in System B. The performance curves have been obtained assuming perfect channel state information in system A and no equalization adaptation error in System B. It can be seen that System A outperforms System B. The performance of System A as compared to System B is expected to get better at lower bit error rates because of the more efficient use of diversity by System A.

Figure 4, where it is seen that an increase in M, in the case of System A, decreases the probability that the PMEPR exceeds an arbitrary threshold. Regarding the complexity of the proposed system, [4] shows that the overall complexity in generating the signal points to be used in evaluating the bit metric is $\mathcal{O}(M^6)$ and the additional Cholesky decomposition that is required for fading channels adds a further complexity of $\mathcal{O}(M^3)$. However, these expressions are misleading for small M, where simulations indicate that the number of computations is greatly reduced through the proposed algorithm.

4 Summary

The performance of the scheme suggested in [1],[2] as a hybrid SC-MC system is studied and compared to an alternate more commonly considered hybrid scheme. It is seen that the performance of the proposed method is superior to the standard scheme under the assumption of perfect channel estimation at the receiver. A modified reduced complexity algorithm for the proposed scheme is presented and results in a large savings in complexity compared to the brute-force approach.

References

[1] D. Goeckel, "Coded Modulation with Non-Standard Signal Sets for Wireless OFDM

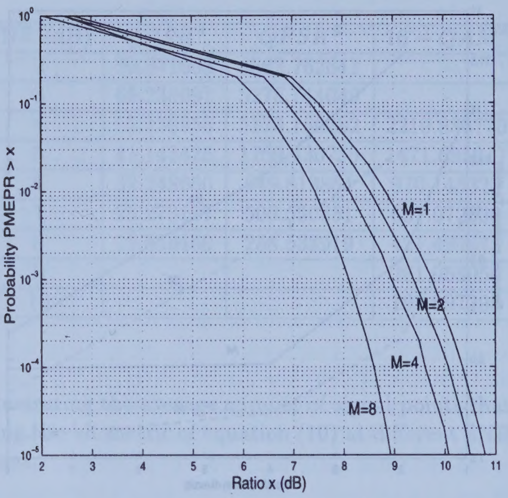


Figure 4: The probability that the PMEPR for a given OFDM symbol exceeds a given level for various signal set dimensions (assuming a total number of carriers of N=16) and $\mathcal{J}$=QPSK. For this calculation, the baseband waveform has been sampled sixteen times between the locations of the IFFT outputs. The number of OFDM symbols simulated for each curve is 5×10^5. Note that M=1 corresponds to a standard OFDM.

Systems," *Proceedings of the 1999 International Conference on Communications*, vol. 2, pp.791-795, Jun. 1999.

[2] D. Goeckel and G. Ananthaswamy, "Increasing Diversity with Non-Standard Signal Sets in Wireless OFDM Systems," *Proceedings of the 1999 IEEE Wireless Communications and Networking Conference*, pp. 20-24, Sept. 1999.

[3] G. Caire, G. Taricco and E. Biglieri, "Bit-Interleaved Coded Modulation," *IEEE Transactions on Information Theory*, vol. 44 , no. 3, pp. 927-946, May 1998.

[4] U. Fincke and M. Phost, "Improved Methods for Calculating Vectors of Short Length in a Lattice, Including a Complexity Analysis," *Math. of Comp.*, vol. 44, no. 170, pp. 463-471, Apr. 1985.

[5] E. Viterbo and J. Boutrous, "A Universal Lattice Decoder for Fading Channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1639-1642.

[6] S. B. Weinstein and P. M. Ebert, "Data Transmission by Frequency-Division Multiplexing Using the Discrete Fourier Transform," *IEEE Transactions on Communication Technology*, vol. 19, no. 5, pp. 628-634, Oct. 1971.

[7] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications,* Third Edition, Prentice Hall, 1996.

[8] G. D. Forney, "Coset Codes - Part I: Introduction and Geometrical Classification," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1123-1151, Sept. 1988.

Performance Analysis of Distributed Dynamic Channel Assignment with Power Control for Cellular Mobile Radio Systems

Zhi Hua Zheng and W. H. Lam, *Senior Member, IEEE*

Dept. of Electrical & Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
e-mail: zhzheng@eee.hku.hk

Abstract

The study of the dynamic channel assignment (DCA) that had been widely carried out was primarily based on simulation results. The corresponding theoretical analysis is seldom. In this paper, the performance of distributed dynamic channel assignment with power control (DDCA/PC) combining both the traffic adaptation and the interference adaptation has been investigated. By employing the reuse factor "*phantom*" N_P, a close form expression to approximate the call blocking probability for the DDCA/PC has been derived. The theoretical results resemble closely to that of the simulation [10]. It has also been shown that the call blocking probability of the DDCA/PC is substantially reduced by more than 350% and 300% compared to that of the DCA and fix channel assignment (FCA) schemes, respectively.

Index Terms----Distributed Dynamic Channel Assignment, Power control, Cellular mobile radio systems

I. INTRODUCTION

The demand in mobile telephone services has seemingly ever increased over the last two decades. Due to the limited frequency spectrum, an efficient channel assignment scheme becomes an important issue in cellular network planning. Both the fixed channel and dynamic channel assignment schemes had been widely studied [5,7,8,9]. The superiority of the dynamic channel assignment (DCA) over the fixed channel assignment in terms of system capacity larger than 30% had been reported in literatures [5,7,8] under the light to medium traffic load. This is due to the fact that each cell in DCA may be allocated with any channel according to the traffic loading and electromagnetic compatibility constraints. In practice, the distributed DCA is preferred as the channel allocation process is distributed among base stations, instead of centralized using the Maximum Packing algorithm [7]. The computation and communication among base stations are distributed amongst base stations and thus reduced. Simulation results had shown [3,5,10,11] that the system capacity for a system employing the distributed DCA equipped with the power has increased substantially by 300%. This paper investigates the distributed DCA techniques which can be further divided into two classes, namely, the Traffic Adaptation DCA (TA-DCA) and the Interference Adaptation DCA (IA-DCA). The TA-DCA scheme performs channel assignments based on the knowledge of the active channels in the neighboring cells and hence it allows the system to manage radio resource according to traffic variations. The IA-DCA scheme performs channel assignment according to the interference in real time. Due to the difficulty in the theoretical analysis for the two schemes, performance evaluation based on simulations had been previously employed and reported in [4,5,9,10]. In this paper, the distributed dynamic channel assignment with power control (DDCA/PC) combining both the traffic adaptation and the interference adaptation are introduced in Section II. A novel analytical methodology which employs the reuse factor "*phantom*" N_P is proposed to analyze the blocking probability in Section III. Section IV describes the analytical results followed by the conclusion in Section V.

II. SYSTEM MODEL

This section describes the system model for the distributed dynamic channel assignment with power control (DDCA/PC) combining both the traffic adaptation and the interference adaptation. The system model considered is based on a mobile radio cellular system having regular hexagonal cells. It employs FDMA or TDMA schemes. Assume that there are L orthogonal available channels in the system which are divided into L/N groups, where N is the system cluster size. Each cell in the cluster has L/N available channels (nominal channels) for the fix channel assignment to allocate channels, but it may also allocate the other X available channels ($X=L-L/N$) under the interference constraints if there is no available nominal channel in the call arising cell. The interference includes the co-channel interference only and the adjacent channel interference is neglected. The DDCA/PC scheme is therefore to employ a cost function to allocate channels followed by the power control in an attempt to reduce the co-channel interference as described in the following subsections. The channel allocation amongst the total available channels is based on the cost function which calculates the interference level so that the minimum one is selected to serve the new arrival call.

A. Cost Function

The cost functions employed in [8,9,10] were slightly different to each other but the prime constraints remain basically the same. They can be collectively expressed in a general expression:

$$F_k=\sum_{i\in I_c}(C_{k_i}q_{k_i})+q_cC_k \tag{1}$$

where F_k is the channel interference cost unit for kth channel, I_c denotes the set of co-channel interference cells about cell C, the notation C_{ki} denotes the binary status of I_c which signifies that $C_{ki}=0$ if channel k is available in the cell i and $C_{ki}=1$ if channel k is currently used in the cell i. The weighting factor q_{ki} is used to reflect the distance between the interfering cell i and the cell C. The notation C_k denote the binary status of cell C which signifies that $C_k=1$ when there is no nominal channels in the cell C, otherwise $C_k=0$. The weighting factor q_c is used to reflect the nominal channel occupied grade. The value of q_c is in proportion to that of q_{ki}, normally $q_c=1$ if the nominal channel is being used in cell C. Therefore, any available channel having minimum value of F_k is to be allocated to a new call arising in cell C.

B. Power Control

The power control is also employed into the channel allocation schemes (DDCA/PC) as follows. Let P^d_{ik} ($P^d_{ik}>0$) denotes *k-th* channel down link effective radiation transmission power from the base station of the cell i and G_{ij} ($G_{ij}>0$) be the radio propagation gain on the path from the base station of cell j to the mobile user in cell i. Suppose that the same channel is used in cell i and reused in cell j with the transmitter power P^d_{jk}, the product $G_{ij}P^d_{jk}$ becomes the amount of co-channel interference to the active user in cell i from cell j. The signal to interference ratio (*SIR*) for the *k-th* channel in the cell *i* can thus be generally written as

$$SIR_i = \gamma_i = \frac{G_{ii} P^d{}_{ik}}{\sum_{j \neq i}^{M} G_{ij} P_{jk}{}^d + V_i} \tag{2}$$

where *M* is the number of co-channel users in a system and V_i is the Additive White Gaussian Noise (AWGN) of the cell *i*. If there exists a power vector $P^d = [P_1{}^d, P^d{}_2, \ldots P^d{}_M]$ which is used to represent the respective ERP power of each co-channel in a system, such that $\gamma_i \geq \gamma$ for $i = 1,2,\ldots M$, where γ is the system determined *SIR* threshold value, the channel *k* is said achievable and hence can be allocated to cell *i*. The distributed power control algorithm to search for a locally optimal power for the new call can thus written as:

$$P^d{}_i(k+1) = min\{\ \eta_i(k) P^d{}_i(k),\ P^d{}_{max}\ \} \tag{3}$$

where $$P^d(0) = P^d{}_{min}, \qquad \eta_i(k) = \gamma / \gamma_i(k) \tag{4}$$

$P^d(0)$ and $P^d(k)$ denote the initial and *k*-th discrete time transmitted power vector, respectively; $P^d{}_{min}$ and $P^d{}_{max}$ are the minimum and maximum transmitted power respectively. The power control is thus to optimize the overall quality for all on-going calls with the same channel. Furthermore, the up-link power control which is not within the scope of this discourse can also be derived in a similar manner.

III. PERFORMANCE ANALYSIS

This section presents the analysis for the call blocking probability of DDCA/PC scheme. Consider a cellular system in which both the mobile received signal strength *S* and the mobile received total interference strength *I* depend on both the user location within the cell and the co-channel users of the other cells. The uplink interference is not considered. It is therefore desirable to allocate channel to call to achieve its maximum system capacity as long as the *S/I* of each individual call is not less than the threshold γ. The DDCA/PC is thus to exercise the power control in order to compromise between the S/I and the system capacity. By adopting the power control into the DDCA which leads to a near-zero variance of *S/I* and ignoring shadow fading, the system capacity C_p for the DDCA/PC can thus be expressed in terms of *S/I* and path loss [3] as

$$C_p = A\ 2^{\frac{-S/I}{1.5\alpha}} \tag{5}$$

where *A* is a constant and α is the path loss exponent. Considering a fix channel assignment (FCA) scheme, denotes the *S/I* gain above the threshold γ for the FCA system as $\Delta G_{FCA} = S/I_{FCA} - \gamma$, where S/I_{FCA} is the median value of the signal to interference ratio. Similarly for the DDCA/PC scheme, denotes the *S/I* gain above the threshold γ as $\Delta G_{DCA} = S/I_{DCA} - \gamma$ where S/I_{DCA} is the median value of the signal to interference ratio for the DCA scheme. The resultant DCA gain in term of *S/I* is given by $\Delta G = \Delta G_{FCA} - \Delta G_{DCA}$, where ΔG in unit of dB. The resultant interference adaptation gain of the DDCA/PC relative to that of the FCA is approximately given by

$$g_{IA} = min\,(\,N,\, 2^{\frac{\Delta G}{1.5\alpha}}\,) \tag{6}$$

where g_{IA} is power ratio, N is the number of cells in a cluster. Suppose that there exists a virtue cluster in the DDCA/PC scheme, namely, the new reuse factor *"phantom"* N_P, which is given by

$$N_p = \frac{N}{g_{IA}} \tag{7}$$

By introducing N_p into the DCCA/PC, the call blocking probability is derived as follows. Let P_i be the probability of having i number of channels occupied. According to the standard *Erlang-B* formula, the stead-state probability distribution P_i is given by

$$P_i = \frac{(N_p\lambda)^i}{i!\,\mu^i} P_0, \qquad 0<i\le L \tag{8}$$

and

$$P_0 = \left(\sum_{k=0}^{L} \frac{(N_p\lambda)^k}{k!\,\mu^k} \right)^{-1} \tag{9}$$

where L is the total number of channels in system, λ is the call arrival rate in per cell following an independent Possion process and $1/\mu$ is the mean call holding time. Obviously, when all the total L channels are occupied in a cluster of size N_p, the call blocking probability is given by P_L. However, in the DDCA/PC schemes, a new call may be blocked even if the number of channels occupied in its reuse cluster N_P is less than L. Let $b(m)$ be the channel occupied probability [2][6] while a channel cannot be assigned purely due to the excessive interference (i.e., S/I < 10*dB*) given m number of channels are occupied in the cluster size of N_p. The $b(m)$ for DDCA/PC is thus approximately given by

$$b(m) = \left(\frac{(1-P_L)D}{(g+1)L/N_P} \right)^{\left\lfloor \frac{L-m}{N_P} \right\rfloor} \tag{10}$$

where
$$D = \frac{\lambda}{\mu}\left(1+\frac{g}{N_p}\right) \tag{11}$$

$$g = 2(N_p-1) \tag{12}$$

In the above expressions, the notation D is the traffic offered to the dynamic channels, the term $[(1-P_L)D]/[(g+1)L/N_p]$ is the traffic carried by an individual channel per cell, the term $\lfloor (L-m)/N_p \rfloor$ is the approximate average number of unused channels per cell in the cluster of size N_P, the notation g is the number of co-channel interfering cells for a given cell with the reuse factor *"phantom"* N_p and the operator $\lfloor x \rfloor$ denotes the integer part of real number x. By applying the channel occupation probability to equation (10), the call arrival rate λ_m given m channels being occupied in a cluster N_P is written as

$$\lambda_m = N_P\,\lambda\,\{1-b(m)\} \tag{13}$$

The stead-state probability distribution of the DDCA/PC given *j* number of channels being occupied in a cluster of N_p is written as

$$P_j = \frac{\prod_{m=0}^{j-1} \lambda_m}{j!\mu^j} P_0 \tag{14}$$

where $$P_0 = \left(1 + \sum_{k=1}^{L} \frac{\prod_{m=0}^{k-1} \lambda_m}{k!\mu^k} \right)^{-1} \tag{15}$$

Thus the system call blocking probability is equal to

$$P_B = \sum_{j=0}^{L} b(j) P_j \tag{16}$$

where *b(j)* and P_j are given in equation (10) and (14), respectively.

IV. NUMERICAL RESULTS

This section presents the performance of the DDCA/PC, FCA and the ordered-DCA (ODCA) channel assignment schemes. For the sake of comparison between the three schemes, we assume N=7, L=70 and $1/\mu$ =120sec. As shown in figure 1 when λ=0.15, the call blocking probabilities for the DDCA/PC ($\Delta G = 10dB$), the FCA and the order-DCA schemes are respectively given by 0.011, 0.49 and 0.50. Observe that the FCA is superior to ordered-DCA when the traffic load is high (i.e. $\lambda \geq 0.2$ in Fig.1). For light to medium traffic load level (i.e. $\lambda < 0.2$ in Fig.1), the ordered-DCA becomes superior instead. Moreover, the DDCA/PC has greatly increased the system capacity and reduced the blocking probability by more than 350% and 300% compared to that of the ODCA and the FCA, respectively. Note that our theoretical results resemble to that of the simulation results obtained from [10]. For instance, the difference between the two results is no more than 8% and 6% for light traffic and high traffic, respectively. Figure 2 depicts the performance of the DDCA/PC algorithms having different values of ΔG. The blocking probabilities P_B decreases with increasing ΔG. An increase of $2dB$ in ΔG, P_B reduces by more than 300% and 100% for light traffic (i.e. $\lambda < 0.2$ in Fig.2) and high traffic (i.e. $\lambda \geq 0.2$ in Fig.2), respectively. The respective system capacity has therefore increased by about 70% and 50%. The increase in ΔG adapts larger *S/I* variance by power control to increase the system capacity.

V. CONCLUSION

As a close form expression for the blocking probabilities of the distributed DCA schemes is formidable, many of the previous research works were based on simulation. This paper presents a close for expression to approximate the call blocking probability for the DDCA/PC algorithms using *"phantom"* N_P. It has been shown that the analytical results resemble to that of simulation [10] having the difference of less than 8%. The analytical results had also shown that the DDCA/PC scheme reduces the blocking probability by more than 300% and 350% compared to that of FCA and ODCA schemes, respectively. An increase of 2dB in ΔG would reduce P_B by more than 300% and 100% for light traffic and high traffic, respectively.

References

[1]D. Kim, K. Chang and S. Kim, "Efficient Distributed Power Control for Cellular Mobile systems ", IEEE *Trans. VT-46, No.2,* May 1997, pp. 313-319

[2]S. S. Kuek and W.C. Wong, "Approximate Analysis of A Dynamic-Channel-Assignment scheme with Handoffs ", IEE *Proc. Commun, Vol.141, No.2,*April 1994, pp.89-92

[3]J. F.Whitehead, " Performance and Capacity of Distributed Dynamic Channel Assignment and Power Control in Shadow Fading", *ICC'93, Geneva*, IEEE *Inter. Conf. Vol.2*, pp.910-914, May,1993

[4]Y. Argyropoulos and S.P.R. Kumar, "Capacity Improvement in Cellular System through Distributed, C/I-Based Power Control", *UPC'1996*, IEEE *Inter. Conf. Vol.1*, pp.164-168, Sep., 1996

[5]F. Borgonovo, A. Capone and A. Molinaro, " Performance of A Dynamic Channel Allocation Technique Based on Signal Strength Measures", *VTC'98, 48th*, IEEE *Vol.2*, pp.1385-1389, May 1998

[6]Wuyi Yue, "Analytical Methods to Calculate the Performance of a Cellular Mobile Radio Communication System with Hybrid Channel Assignment ", IEEE *Trans. VT-40, No.2*, pp. 453-460, May 1991

[7]L. J. C. Jr., G. J. Foschini, Chih-Lin I. Zoran M., "Call Blocking Performance of Distributed Algorithms for Dynamic channel Allocation in Microcells", IEEE *Trans. on Commun. Vol 42 No.8,* pp.2600-2607, Aug., 1994

[8]E. Del Re, R. Fantacci, G. Giambene, "Call Blocking Performance for Dynamic Channel Allocation Technique in Mobile Satellite Systems", IEE *Proc.Commun., Vol. 143, No.5,* pp.289-296, Oct., 1996

[9]C. Carciofi, G. Falciasecca, M. Frullone, N. P. Magnani, V. Palestini, " Traffic-adaptation and interference adaptation resource allocation scheme in GSM and DCS1800 systems under realistic traffic and propagation conditions: a case study" IEEE 46th, *Mobile Technology for the Human Race,* pp.497-501, 1996

[10]Shaoji Ni, " Distributed Channel allocation Algorithm with power Control", PIMRC'97, IEEE *International Symposium on Vol.2,* pp.406-410, 1997

[11]S. A. Grandhi, Roy D. Yate and David J. Goodman, " Resource Allocation for Cellular Radio Systems ", IEEE *Tran. VT-46*, No.3, Aug. 1997, pp581-587

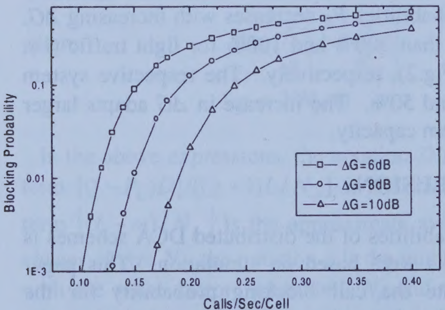


Fig.1 The call blocking probability of DDCA/PC schemes having different values of *ΔG*

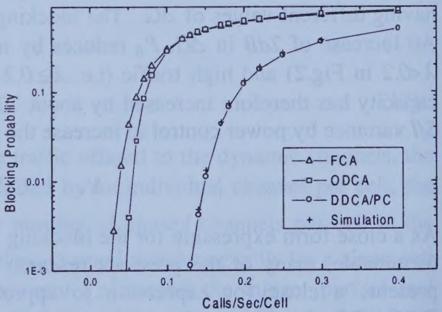


Fig.2 The call blocking probability of FCA, ODCA and DDCA/PC schemes

Optimal Distributed Power Control in Cellular Wireless Systems

Lijun Qian and Zoran Gajic
Department of Electrical and Computer Engineering
Rutgers—The State University of New Jersey
94 Brett Road, Piscataway, NJ 08854-8058
ljqian, gajic@ece.rutgers.edu

Abstract

In this paper, an optimal fast closed-loop SIR-based power control scheme is proposed for CDMA systems. The state-of-the-art of this power control algorithm is that the power evolution is directly driven by the error between the actual SIR and the desired SIR. The amount of power change is proportional to the error, and the optimal gain is chosen to achieve the minimum error at the next time instant. Assuming perfect estimation and prediction, the error between the desired SIR and the real value of SIR goes to zero in one step. In practical situations, where the estimator need several (3-4) steps to make accurate estimation and prediction, the error between the desired SIR and the real value of SIR goes to zero in several (4-5) steps. The proposed controller is an estimator based controller. The performance of the estimator will have direct impact on the performance of the power control scheme. Simulations of a CDMA system are performed to demonstrate effectiveness of the proposed power control scheme. The simulation results, in all cases, confirm theoretical analysis. These results indicate that the proposed optimal power control scheme is a very promising solution to the distributed fast closed-loop power control problem in wireless systems, both theoretically and practically.

I. REVIEW OF SIR BASED POWER CONTROL IN CELLULAR WIRELESS SYSTEMS

In a cellular wireless network, certain quality-of-service (QoS) should be maintained for all active users in the network. A quantity that measures the user's provided QoS is the signal-to-interference ratio (SIR). The initial work on SIR-based power control schemes for narrowband systems has been done by Zander [1]-[3].

Power control schemes can be centralized [1]-[3] or distributed [2],[4],[6] according to the nature of computations. A centralized controller has information (e.g., all the link gains) for each user and it decides control actions for all users. On the other hand, a distributed controller only uses local information to make a control action for a user. For example, [2],[4],[6] only use information about user's own link gain and/or SIR to make a decision about their transmitted power.

The algorithm convergence has to be studied carefully when a distributed power control scheme is used. The distributed power control algorithm by Foschini [6] was shown to converge either synchronously [6] or asynchronously [9]. The basic model in [6] is relaxed in [9] to allow asynchronism. Propagation delay is also considered in [9]. A framework for convergence of the generalized uplink power control was proposed by Yates [10].

In recent years, there have been increasing interests in applying code-division-multiple-access (CDMA) techniques to cellular wireless networks. Unlike the narrowband systems, users in a CDMA system share the same frequency all the time by using a specific spread spectrum pseudonoise code for each user. The system performance mainly depends on so-called multiple access interference (MAI), which is caused by the cross-correlations between the desired signal and other signals.

Open-loop power control has been employed to combat the path loss and shadow fading. The average power control in [10],[22] keeps the local mean received power to a constant value, which mitigate the effect of shadowing and near-far problems. In IS-95B [20] and CDMA2000 [21], fast closed-loop power control is proposed to fight medium to fast fading.

In [11]-[12], it is shown that when power control rate is higher, it can partially accomodate the effect of fast fading.

Herdtner and Chong [9] considered a simple distributed power control scheme in which the convergence condition was given. Wu [13] extended Zander's problem formulation to CDMA systems by reordering the number of users in different cells. Song et.al. [14]-[15] and Gunnarsson et.al. [16]-[17] considered up/down power control and closed-loop power control with other nonlinear elements within the framework of nonlinear control systems. [14]-[15] gave guidelines for choosing the appropriate power control step size in IS-95. In [16]-[17], PID controllers were designed to overcome the effects of time-delay in the feedback loop.

Filtering techniques have been also used to design power control schemes. Leung [18] proposed a power control scheme for TDMA data service based on the Kalman filtering technique. In [31], a linear prediction of received power is made at the base station to predict the power control bit one step ahead. Both approaches assume that the interference is Gaussian.

In this paper, an optimal fast closed-loop SIR-based uplink power control scheme is proposed for CDMA systems. In Section 2, a proportional controller is proposed and the optimal power control gain is derived by unconstrained optimization. In Section 3, the constraints are imposed on the maximum and minimum values of transmitter power. By applying the Kuhn-Tucker conditions, we derive the optimal gain for the proportional controller. Using Bellman's optimality principle, the power control scheme achieves optimality in the whole time interval of interest. The proposed controller is an estimator based controller. The performance of the estimator will have direct impact on the performance of the power control scheme. Section 4 gives our choice of the estimator. In contrast to the well-known Kalman filter, which assumes that system and measurement disturbances are white Gaussian stochastic processes and requires the knowledge of their statistics, the H_∞ filter, used in the proposed scheme, facilitates the system filtering regardless of the nature of the system and measurement disturbances. It meets our estimation requirement since the fluctuation of the interference is not Gaussian and its statistics is not known in general [18]. Simulations of a CDMA system are performed to demonstrate the effectiveness of the proposed power control scheme. The results are shown in Section 5.

II. OPTIMAL SIR-BASED DISTRIBUTED POWER CONTROL BY UNCONSTRAINED OPTIMIZATION

In this paper, only uplink (mobile-to-base) power control is considered. Suppose that there are N active users in the system. The Signal-to-Interference Ratio (SIR) in a CDMA system is defined as

Definition 1: The SIR of an active link from mobile station i to base station n in a wireless system is defined as

$$\gamma_{ni} = \frac{Lh_{ni}p_i}{\sum_{j\neq i} h_{nj}p_j + \sigma^2} \tag{1}$$

where h_{ni} is the link gain from mobile station i to base station n. We assume that L is the processing gain in spread spectrum wireless systems, e.g., in CDMA 2000, L=64 or 128 or 256. Let us denote the denominator by I_i, which represents the received interference, then the SIR can be rewritten as

$$\gamma_{ni} = \frac{Lh_{ni}p_i}{I_i} \tag{2}$$

Assuming that a mobile only transmits to one base station during the time of power control, then γ_{ni} can be simplified as γ_i. Let $\delta_i = \frac{Lh_{ni}}{I_i}$ denote the channel variation [11]. δ_i will be estimated and predicted in the proposed power control scheme.

Given that a mobile station and its base station communicate directly to each other, we assume that h_{ni} can be estimated accurately. We also assume that the received interference I_i can be measured at each time step. The received interference is the difference between the total received power and the desired signal power [10].

The main idea of the proposed algorithm is to make the transmission power proportional to the error between the actual SIR and the desired SIR. Define the transmission power change from time step k to $k+1$ as

$$\Delta p_i(k+1) = p_i(k+1) - p_i(k) \tag{3}$$

where $p_i(k)$ and $p_i(k+1)$ are the transmission powers from mobile i to the base station at the time instants k and $k+1$, respectively. Let the error between the actual SIR ($\gamma_i(k)$) and the desired SIR (γ_i^{tar}) be denoted by $e_i(k)$ for mobile i at the time instant k,

$$e_i(k) = \gamma_i^{tar} - \gamma_i(k) \tag{4}$$

We propose the following power control algorithm

$$\Delta p_i(k+1) = \alpha_i(k) e_i(k) \quad \Longrightarrow \quad p_i(k+1) = p_i(k) + \alpha_i(k)(\gamma_i^{tar} - \gamma_i(k)) \tag{5}$$

where $\alpha_i(k)$ is the gain to be determined through the optimization procedure in which the square of the error is minimized (theoretically reduced to zero in each discrete-time instant k). The main result of this paper is presented in the following theorem.

Theorem 1: Define the performance criterion at the time instant k as

$$J_i(k) = \min_{\alpha_i(k)} (e_i(k+1))^2 \;, \quad i = 1, 2, ..., N. \tag{6}$$

The optimal gains of the power control algorithm (5) that minimize $J_i(k)$ at every time instant k are given by

$$\alpha_i^{opt}(k) = \begin{cases} \frac{1}{\delta_i(k)}(1 - \frac{\gamma_i^{tar}}{e_i(k)}) + \frac{1}{\delta_i(k+1)} \frac{\gamma_i^{tar}}{e_i(k)} & if \quad e_i(k) \neq 0 \\ 0 & if \quad e_i(k) = 0 \end{cases} \tag{7}$$

where $\quad \delta_i(k) = \frac{L h_{ni}(k)}{I_i(k)} > 0, \quad \delta_i(k+1) = \frac{L h_{ni}(k+1)}{I_i(k+1)} > 0$
and $\quad h_{ni}(k), h_{ni}(k+1), I_i(k), I_i(k+1) > 0, \ \forall k.$
The minimal value of the performance criterion is zero when the optimal gain $\alpha_i^{opt}(k)$ is applied.

Proof: Firstly, we derive the error dynamic equation using (1)-(5)

$$e_i(k+1) = (-\delta_i(k+1)\alpha_i(k) + \frac{\delta_i(k+1)}{\delta_i(k)}) e_i(k) + (1 - \frac{\delta_i(k+1)}{\delta_i(k)}) \gamma_i^{tar} \tag{8}$$

Secondly, we compute the square of the error

$$\begin{aligned} (e_i(k+1))^2 &= (-\delta_i(k+1)\alpha_i(k) + \frac{\delta_i(k+1)}{\delta_i(k)})^2 (e_i(k))^2 + (1 - \frac{\delta_i(k+1)}{\delta_i(k)})^2 (\gamma_i^{tar})^2 \\ &+ \ 2(-\delta_i(k+1)\alpha_i(k) + \frac{\delta_i(k+1)}{\delta_i(k)}) e_i(k) (1 - \frac{\delta_i(k+1)}{\delta_i(k)}) \gamma_i^{tar} \end{aligned} \tag{9}$$

From the necessary conditions for optimality at the time instant k

$$\frac{\partial J_i(k)}{\partial \alpha_i(k)} = \frac{\partial (e_i(k+1))^2}{\partial \alpha_i(k)} = 0 \tag{10}$$

we have

$$\alpha_i^{opt}(k) = \frac{1}{\delta_i(k)}\left(1 - \frac{\gamma_i^{tar}}{e_i(k)}\right) + \frac{1}{\delta_i(k+1)}\frac{\gamma_i^{tar}}{e_i(k)} \tag{11}$$

Taking the second derivative we obtain

$$\frac{\partial^2 J_i(k)}{\partial \alpha_i^2(k)} = 2(\delta_i(k+1)e_i(k))^2 > 0 \tag{12}$$

which indicates that the resulted $J_i(k)$ by using $\alpha_i^{opt}(k)$ is at minimum and establishes the sufficiency part of the theorem. Another obvious proof of sufficiency follows from the fact that $J_i(k) \geq 0$ and $J_i(\alpha_i^{opt}(k)) = 0$. Using Bellman's optimality principle [29], we achieve optimality in the whole time interval of interest by applying this power control algorithm at each time step. In other words, the controller gains obtained in (7) are optimal for all k. ■

The block diagram of the proposed distributed power control algorithm is shown in Figure 1. The state-of-the-art of this power control algorithm is that the power evolution is directly driven by the error between SIR and the desired SIR. The amount of power change is proportional to the error, and the gain is chosen to achieve the minimum error at the next time instant $k+1$.

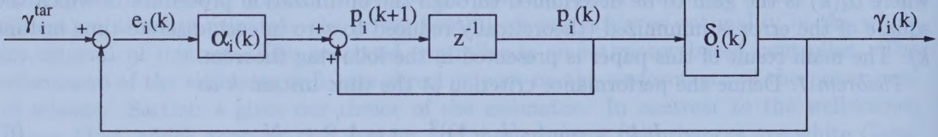


Figure 1. Optimal SIR-based distributed power control of cellular wireless systems

In traditional CDMA systems, e.g., IS-95, power control was used to combat path loss and shadow fading. Rayleigh fading was compensated by advanced modulation and coding schemes. On the other hand, in 3-G wireless systems, e.g., CDMA-2000, UMTS, the closed-loop power control is faster than before. For example, in IMT-2000 proposal [21], the frequency of closed-loop power control is 1600 Hz. It should be fast enough to compensate Rayleigh fading. Traditional approaches using power iterations or the implemented up/down power control schemes are not appropriate to compensate Rayleigh fading, because both schemes requires rather long time to converge. Usually, they assume that the link gains and fading conditions are not changed much during the whole power control period.

An important assumption behind all the power iterative algorithms reviewed in Section 1 is that power updates occur so quickly that the link gains can be assumed to be fixed. Our power control algorithm is *not* an iterative algorithm. The assumption that the link gains do not change during power evolution is *not* required in the proposed algorithm. On the other hand, the proposed power control algorithm is suitable for such purpose due to the immediate decision of the optimal power allocation of the next time step based on measurements and predictions. With a good estimator (for example, discrete-time H_∞ filter), all the effects including path loss, shadow fading and Rayleigh fading, will be taken into consideration. The error between the actual SIR and the desired SIR is driven to zero (theoretically) in every single step.

The block diagram of the proposed power control scheme with an estimator is shown in Figure 2, where $f()$ is defined as

$$f(x_1, x_2) = \begin{cases} \frac{1}{x_1}\left(1 - \frac{\gamma_i^{tar}}{e_i(k)}\right) + \frac{1}{x_2}\frac{\gamma_i^{tar}}{e_i(k)} & if \quad e_i(k) \neq 0 \\ 0 & if \quad e_i(k) = 0 \end{cases} \tag{13}$$

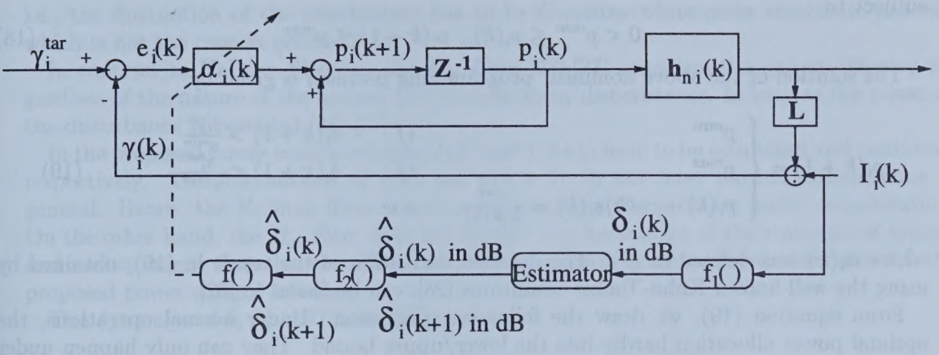


Figure 2. The optimal distributed power control system with an estimator

$f_1(x) = 10log_{10}(x)$ and $f_2(x) = 10^{x/10}$ are converting functions between W and dBW.

It is noticeable that the proposed controller is an estimator based controller. The idea of quick online estimation of link quality was mentioned in [19]. It pointed out that any adaptive power control scheme need quick detection of the change of link quality. The performance of the estimator will have a direct impact on the performance of the power control scheme. Let $\hat{\delta}_i(k)$ and $\hat{\delta}_i(k+1)$ denote the estimated/predicted value of $\delta_i(k)$ and $\delta_i(k+1)$, respectively. Then the gain applied to the power control algorithm is

$$\hat{\alpha}_i(k) = \begin{cases} \frac{1}{\hat{\delta}_i(k)}(1 - \frac{\gamma_i^{tar}}{e_i(k)}) + \frac{1}{\hat{\delta}_i(k+1)}\frac{\gamma_i^{tar}}{e_i(k)} & if \quad e_i(k) \neq 0 \\ 0 & if \quad e_i(k) = 0 \end{cases} \tag{14}$$

The error between the actual SIR and the desired SIR at the next time instant $k+1$ is

$$\begin{aligned} e_i(k+1) &= \delta_i(k+1)[(\frac{1}{\delta_i(k)} - \frac{1}{\hat{\delta}_i(k)})(e_i(k) - \gamma_i^{tar}) + (\frac{1}{\delta_i(k+1)} - \frac{1}{\hat{\delta}_i(k+1)})\gamma_i^{tar}] \\ &= p_i(k+1)(\hat{\delta}_i(k+1) - \delta_i(k+1)) \end{aligned} \tag{15}$$

It is obvious that if the estimated/predicted values of $\delta_i(k)$ and $\delta_i(k+1)$ are close to the real values, then $e_i(k+1) \approx 0$. The power evolution is given by

$$p_i(k+1) = p_i(k) + \hat{\alpha}_i(k)e_i(k) = \frac{\gamma_i^{tar}}{\hat{\delta}_i(k+1)} \tag{16}$$

III. OPTIMAL SIR-BASED DISTRIBUTED POWER CONTROL BY CONSTRAINED OPTIMIZATION

In the previous section, we have performed unconstrained optimization to get the optimal gain for the controller. However, in real wireless systems, there are always constrains on the power that the mobile can transmit, namely, the highest transmitted power allowed and the lowest transmitted power to maintain connections. In this section, we will derive the optimal gain for our controller by using constrained optimization.

Suppose that the highest transmitted power allowed is p^{max} and that the lowest transmitted power allowed is p^{min}. We formulate the following constrained optimization problem

$$J_i(k) = \min_{\alpha_i(k)}(e_i(k+1))^2 \tag{17}$$

subject to

$$0 < p^{min} \leq p_i(k)\ ,\ p_i(k+1) \leq p^{max} \tag{18}$$

The solution of the above nonlinear programming problem is given by

$$p_i(k+1) = \begin{cases} p^{min} & if \quad \delta_i(k+1) > \frac{\gamma_i^{tar}}{p^{min}} \\ p^{max} & if \quad \delta_i(k+1) < \frac{\gamma_i^{tar}}{p^{max}} \\ p_i(k) + \alpha_i(k)e_i(k) = \frac{\gamma_i^{tar}}{\delta_i(k+1)} & o.w. \end{cases} \tag{19}$$

where $\alpha_i(k)$ was defined in (7). The detailed derivation of the result in (19), obtained by using the well-known Kuhn-Tucker conditions [28], can be found in [32].

From equation (19), we draw the following conclusion: Under normal operations, the optimal power allocation hardly hits the lower/upper bound. They can only happen under extreme conditions, namely, when the desired SIR is either set too low $\delta_i(k+1)p^{min} > \gamma_i^{tar}$ or too high $\delta_i(k+1)p^{max} < \gamma_i^{tar}$. These extreme conditions can happen in the proposed power control scheme when γ_i^{tar} is not set properly and/or when measurements and/or predictions are far from the real values. These extreme conditions provide a mean to evaluate the quality of the H_∞ filter (or any other estimator to be used) when γ_i^{tar} is set properly. It can be stated as

if $\quad \hat{\delta}_i(k+1) < \frac{\gamma_i^{tar}}{p^{max}}$, or if $\quad \hat{\delta}_i(k+1) > \frac{\gamma_i^{tar}}{p^{min}}$, then the predictions are bad.

However, we should keep in mind that the performance of the H_∞ filter does not depend on the value of γ^{tar}. It mainly depends on the choice of the optimization parameters of the H_∞ filter. These parameters will be introduced in Section 4.

Based on the above discussions, the modified power control algorithm, which can be implemented in existing wireless systems, is proposed as follows:

$$p_i(k+1) = \begin{cases} p^{min} & if \quad \hat{\delta}_i(k+1) > \frac{\gamma_i^{tar}}{p^{min}} \\ p^{max} & if \quad \hat{\delta}_i(k+1) < \frac{\gamma_i^{tar}}{p^{max}} \\ p_i(k) + \hat{\alpha}_i(k)e_i(k) = \frac{\gamma_i^{tar}}{\hat{\delta}_i(k+1)} & o.w. \end{cases} \tag{20}$$

where $\hat{\alpha}_i(k)$ is defined in (14). Formula (20) is consistent with the Pontriagin's minimum principle which stated that in constrained optimization, optimal solutions lie on the boundary [29].

The block diagram of the complete power control system composed of N links with estimators can be found in [32], where the effect of the correlation in the received power from interfering users is taken into account.

In order to compute $\hat{\alpha}_i(k)$, we need a 'good' estimator/predictor. The H_∞ filter is a very promising candidate. We will discuss this issue in the next section.

IV. H_∞ FILTER AS AN OPTIMAL ESTIMATOR/PREDICTOR

The problem of filtering involves estimation of system states using past measurements. The development of efficient linear estimators, (e.g., the Kalman filter), has been based mainly on the minimization of the L_2 -norm of the corresponding estimation error. This type of estimation assumes that the message generating process has a known dynamics and that the exogenous inputs have known statistical properties. The well-known Kalman filter offers optimal filtering algorithm when the system model parameters are known and the system and measurement noise is represented by white Gaussian processes with known statistics (power spectral density). In [18], Leung applied the Kalman filter to predict the interference in the immediate future. Its usage is constrained by the requirement of the Kalman filter, i.e., the fluctuation of the interference has to be Gaussian white noise stochastic process, which is not the case in general [18].

In contrast to Kalman filtering, the H_∞ filter [23]-[27] considers the system filtering regardless of the nature of the system and measurement disturbances, as long as the power of the disturbance is bounded [26]-[27].

In the proposed power control scheme, $\delta(k)$ and $\delta(k+1)$ have to be estimated and predicted, respectively. The fluctuations of $\delta(k)$ and $\delta(k+1)$ do not have normal distributions in general. Hence, the Kalman filter is not appropriate for the problem under consideration. On the other hand, the H_∞ filter does not require any knowledge of the statistics of system and measurement disturbances. This makes the H_∞ filter an aprropriate estimator for the proposed power control scheme.

The dynamics of δ_i (in dB) is described by

$$\delta_i(k+1) = \delta_i(k) + w_i(k) \tag{21}$$

where $w_i(k)$ represents the process noise (disturbance). Let $y_i(k)$ be the measurement of $\delta_i(k)$.

$$y_i(k) = \delta_i(k) + v_i(k) \tag{22}$$

where $v_i(k)$ is the measurement noise. Due to the scalar nature of the problem, it is obvious that the system controllability and observability conditions are met [33]. These conditions are needed for the existence of the appropriate solution of the algebraic Riccati equation [34].

The H_∞ filter (suboptimal) solves the following optimization problem

$$\sup_{Q_i, W_i, V_i} \frac{\sum_{k=0}^{N-1} ||\delta_i(k) - \hat{\delta}_i(k)||^2_{Qi}}{||\delta_i(0) - \hat{\delta}_i(0)||^2_{P^{-1}(0)} + \sum_{k=0}^{N-1} ||w(k)||^2_{W_i^{-1}} + ||v(k)||^2_{V_i^{-1}}} < \frac{1}{\lambda} \tag{23}$$

where Q_i, W_i, V_i are positive weighting parameters chosen by the designer, and λ is a prescribed level of noise attenuation [26]-[27]. The discrete-time H_∞ filter is given by

$$\hat{\delta}_i(k+1) = \hat{\delta}_i(k) + K_i(k)(y_i(k) - \hat{\delta}_i(k)) \tag{24}$$

where $K_i(k)$ is the optimal gain of the H_∞ filter. It is computed from the following equation

$$K_i(k) = P_i(k)(I - \lambda Q_i P_i(k) + V_i^{-1}(k) P_i(k))^{-1} V_i^{-1}(k) \tag{25}$$

Since in our problem $P_i(k)$, Q_i and V_i are all scalars, the gain can be found as

$$K_i(k) = \frac{P_i(k)}{(1 - \lambda Q_i P_i(k) + \frac{P_i(k)}{V_i})V_i} \tag{26}$$

and $P_i(k)$ is the unique positive solution of the following scalar difference Riccati equation

$$P_i(k+1) = P_i(k)(I - \lambda Q_i P_i(k) + V_i^{-1}(k) P_i(k))^{-1} + W_i \tag{27}$$

Again, since $P_i(k)$, Q_i, W_i and V_i are all scalars, (27) can be written as

$$P_i(k+1) = \frac{P_i(k)}{1 - \lambda Q_i P_i(k) + \frac{P_i(k)}{V_i}} + W_i \tag{28}$$

The initial condition of the system is $\delta_i(0) = \delta_{i0}$. The initial condition of the above H_∞ filter is $P_i(0) = P_{i0}$. These initial conditions are arbitrary positive quantities. It can be shown that the steady-state value of $P_i(k)$ satisfies

$$P_{i,ss} = \frac{W_i(\frac{1}{V_i} - \lambda Q_i) \pm \sqrt{(W_i(\frac{1}{V_i} - \lambda Q_i))^2 + 4W_i(\frac{1}{V_i} - \lambda Q_i)}}{2(\frac{1}{V_i} - \lambda Q_i)} \tag{29}$$

Only the positive one is the steady-state solution of $P_i(k)$. It is obvious that the choice of the optimization parameters λ, Q_i, W_i and V_i has an impact on the filter performance. The block diagram of our H_∞ filter is shown in Figure 3.

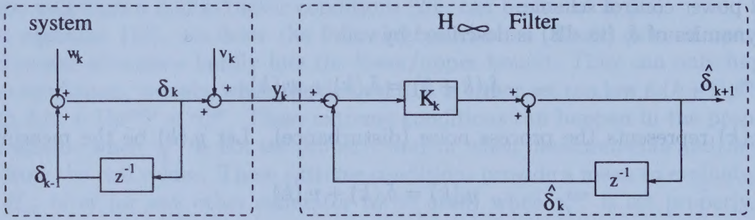


Figure 3. The structure of the discrete-time H_∞ filter

The performance of this discrete-time H_∞ filter will be tested and discussed in the next section.

V. SIMULATION RESULTS AND ANALYSIS

In this section, simulations of a CDMA system are performed to demonstrate effectiveness of the proposed power control scheme. Comparison between the proposed optimal power control scheme and other existing power control schemes clearly shows the advantages of the proposed scheme.

A CDMA system with 7 hexagonal cells and 16 users per cell is considered. The operating frequency is 1.9 GHz and the bandwidth of each channel is assumed to be 1.23 MHz, which is in accordance to the CDMA 2000 standard [21]. Data rate is set at 9600 bps and the processing gain is set at 128 (21 dB). The target SIR is 7 dB, which corresponds to the bit error rate (BER) being less than 10^{-3}. Note that the SIR, as defined in this paper, includes the processing gain, which is denoted by $\frac{E_b}{N_0}$ in the standard IS-95B [20]/CDMA 2000 [21], where E_b is the energy per information bit and N_0 is the interference power spectral density.

In simulation we further make the following assumptions:

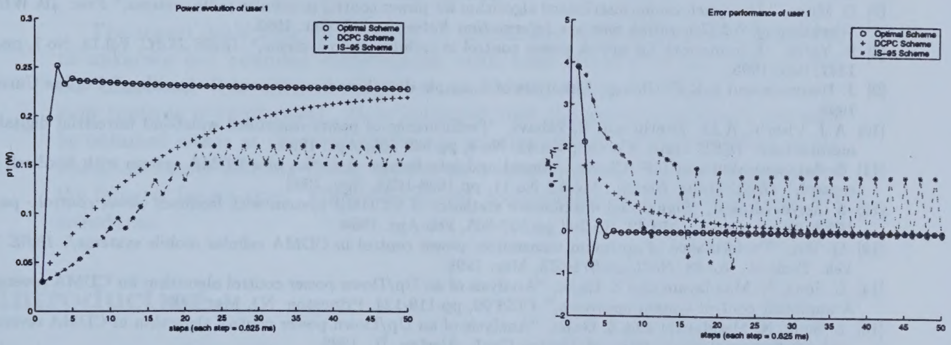
1. The effects of antenna directivity, voice activity factor, and soft handoff are ignored.
2. The minimum and maximum power can be transmitted by a mobile (IS-95B[20]) are $p^{min} = 8$ dBm (6.3 mW) and $p^{max} = 33$ dBm (2 W), respectively.
3. The background noise power is 0.05 mW.
4. The transmitted power is updated periodically, every 0.625 msec, which corresponds to 1,600 Hz fast closed-loop power control frequency proposed in IMT-2000.
5. The antenna of the base station and all mobiles are omnidirectional.
6. Mobiles are assumed to be uniformly distributed in a cell.
7. It is assumed that the link gains have the following form

$$h_{ni}(k) = d_{ni}^{-4}(k) A_{ni}(k) B_{ni}(k) \tag{30}$$

where $d_{ni}(k)$ is the distance from the ith mobile to the nth base station at time instant k, A_{ni} is a log-normal distributed stochastic process, and B_{ni} is contributed by Rayleigh fading [30].

8. It is assumed that the cell diameter is 2 km. $d_{ni}(k)$ is a 2-D uniformly distributed random variable, [30].
9. It is assumed that the standard deviation of A_{ni} is 8, [30].
10. The speed of the mobiles are ranging from 0.5 mph to 60 mph. The maximum Doppler frequency is 80 Hz, which correspond to 60 mph mobile speed. $B_{ni}(k)$ is simulated using a filtered Gaussian noise model, [30].
11. The initial mobile transmitted power is taken as $p_0 = p^{min} + \eta(p^{max} - p^{min})$, where $\eta \in [0, 1]$ is a uniformly distributed random number.
12. The mean values of the link gains are used in the simulation since only the deterministic case is considered in this paper.

The optimization parameters of the H_∞ filter are $\lambda = 0.05$, $Q_i = 20$, $W_i = 1$ and $V_i = 0.1$ for all mobiles. In addition, δ_0 and P_0 are positive random numbers using random number generators.



4(a) 4(b)

Figure 4. Comparison between the optimal power control scheme, DCPC and IS-95:
(a). Power evolution
(b). Error performance

We consider three cases, based on the range of the desired SIR, namely, when γ^{tar} is feasible, when γ^{tar} is set too high, and when γ^{tar} is set too low. The simulation results can be found in [32]. In all three cases the simulation results fit theoretical analysis (Section 2) very well. These results indicate that the optimal power control scheme is a very promising solution of the distributed fast closed-loop power control problem in wireless systems, both theoretically and practically.

In the current IS-95 systems, the power control scheme is up/down hard decision type [20]. It can not achieve zero SIR error, due to persistent oscillations around zero [14].

A widely known and accepted power control scheme is called the Distributed Constrained Power Control (DCPC), and is given by [5]

$$p_i(k+1) = min\{\frac{\gamma_i^{tar}}{\gamma_i(k)}p_i(k),\ p^{max}\} \tag{31}$$

The performances of the proposed optimal power control scheme, DCPC, and IS-95 are compared in Figure 4. When the channel is stationary (the mean values of the channels are constants), the optimal scheme will save at least 2 frames, that would be lost if the other two power control schemes were used.

When channel changes frequently, the estimator used will have good tracking of the channel quality and the proposed optimal power control scheme will also provide accurate values for power adjustments almost instantly. On the other hand, DCPC and up/down scheme in IS-95 will not converge fast enough and will not provide appropriate power changes.

REFERENCES

[1] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Trans. Veh. Technol.*, Vol.41, pp.57-62, Feb. 1992.

[2] J. Zander, "Distributed cochannel interference control in cellular radio systems," *IEEE Trans. Veh. Technol.*, Vol.41, pp.305-311, Aug. 1992.

[3] S. Grandhi, R. Vijayan, D. Goodman and J. Zander, "Centralized power control in cellular radio systems," *IEEE Trans. Veh. Technol.*, Vol.42, No.4, pp.466-468, Nov. 1993.

[4] S. Grandhi, R. Vijayan and D. Goodman, "Distributed power control in cellular radio systems," *IEEE Trans. Comm.*, Vol.42, No.2/3/4, pp.226-228, Feb/Mar/Apr. 1994.

[5] S. Grandhi, J. Zander and R. Yates, "Constrained power control," *Wireless Personal Communications*, Vol.1, pp.257-270, 1995.

[6] G.J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence," *IEEE Trans. Veh. Technol.*, Vol.42, No.4, pp.641-646, Nov. 1993.

[7] D. Mitra, "An asynchronous distributed algorithm for power control in cellular radio systems," *Proc. 4th WINLAB Workshop of 3rd Generation wireless Information Networks*, NJ, Oct. 1993.

[8] R. Yates, "A framework for uplink power control in cellular radio systems," *IEEE JSAC*, Vol.13, No.7, pp.1341-1347, Sep. 1995.

[9] J. Herdtner and E.K.P. Chong, "Analysis of a simple distributed power control algorithm," Purdue University, 1998.

[10] A.J. Viterbi, A.M. Viterbi and E. Zehavi, "Performance of power-controlled wideband terrestrial digital communication," *IEEE Trans. Comm.*, Vol.41, No.4, pp.559-569, Apr. 1993.

[11] S. Ariyavisitakul and L.F. Chang, "Signal and interference statistics of a CDMA system with feedback power control," *IEEE Trans. Comm.*, Vol.41, No.11, pp.1626-1634, Nov. 1993.

[12] S. Ariyavisitakul, "Signal and interference statistics of a CDMA system with feedback power control - part II," *IEEE Trans. Comm.*, Vol.42, No.2-4, pp.597-605, Feb-Apr. 1994.

[13] Q. Wu, "Performance of optimum transmitter power control in CDMA cellular mobile systems," *IEEE Trans. Veh. Technol.*, Vol.48, No.2, pp.571-575, Mar. 1999.

[14] L. Song, N. Mandayam and Z. Gajic, "Analysis of an Up/Down power control algorithm for CDMA reverse link: A nonlinear control system approach," CISS'99, pp.119-124, Princeton, NJ, Mar.1999.

[15] L. Song, N. Mandayam and Z. Gajic, "Analysis of an Up/Down power control algorithm in CDMA reverse link under fading condition," Proc. of Alerton Conf., Alerton, IL, 1999.

[16] F. Gunnarsson, F. Gustafsson and J. Blom, "Improved performance using nonlinear components in power control algorithms," VTC 99.

[17] F. Gunnarsson, F. Gustafsson and J. Blom, "Pole placement design of power control algorithms," VTC 99.

[18] K.K. Leung, "Power control by Kalman filter techniques for broadband wireless networks," IEEE Infocom'99, NY, Mar.1999.

[19] N. Bambos, "Toward power-sensitive network architectures in wireless communications: concepts, issues and design aspects," *IEEE Personal Communications*, vol.5, pp.50-59, 1998.

[20] TIA/EIA Interim Standard-95B: Mobile station-base station compatibility standard for dual-mode wideband spread spectrum cellular system, TIA, 1998.

[21] ITU, RTT 45.5, CDMA 2000 standard.

[22] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver and C.E. Wheatley, "On the capacity of a cellular CDMA system," *IEEE Trans. Veh. Technol.*, Vol.40, No.2, pp.303-312, May. 1991.

[23] K. Zhou, J.C. Doyle and K. Glover, *Robust and Optimal Control*, Prentice-Hall, 1995.

[24] D.S. Bernstein and W.M. Haddad, "Steady-State Kalman Filtering with an H_∞ Error Bound," *Systems and Control Letters*, Vol.12, pp.9-16, 1989.

[25] K.M. Nagpal and P.P. Khargonekar, "Filtering and Smoothing in an H_∞ Setting," *IEEE Tran. on AC*, Vol.36, No.2, pp.152-166, Feb. 1991.

[26] X. Shen and L. Deng, "Game Theory Approach to Discrete H_∞ Filter," *IEEE Tran. on Signal Processing*, Vol.45, No.4, pp.1092-1095, Apr. 1997.

[27] M. Grimble and A. Sayed, "Solution of the H_∞ optimal linear filtering problem for discrete-time systems," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.38, pp.1092-1104, July 1990.

[28] M.S. Bazaraa and C.M. Shetty, *Nonlinear Programming - Theory and Algorithms*, John Wiley & Sons, 1979.

[29] D.E. Kirk, *Optimal Control Theory : An Introduction*, Prentice-Hall, 1970.

[30] G. Stuber, *Principles of Mobile Communication*, Kluwer Academic Publishers, 1996.

[31] S. Choe, T. Chulajata, H. Kwon, B. Koh and S. Hong, "Linear prediction at base station for closed loop power control," VTC 99.

[32] L. Qian, Optimal Distributed Power Control in Cellular Wireless Systems, *Ph.D. Dissertation*, Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, May, 2000.

[33] T. Kalaith, *Linear Systems*, Prentice-Hall, 1980.

[34] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, John Wiley & Sons, 1972.

REACHABILITY UNDER DISTURBANCES*

A.B.Kurzhanski and P.Varaiya

University of California at Berkeley, CA 94720-1770,
Electrical Eng.@ Comput.Sci, Cory Hall,
e-mail: kurzhans@eecs.berkeley.edu varaiya@eecs.berkeley.edu

Abstract

The report deals with the problem of reachability for linear systems subjected to unknown but bounded disturbances, with hard bounds on the controls and the unknowns. The difference between reachability under closed-loop and open-loop controls is emphasized. It is indicated that the closed-loop reach set may be obtained either as the level set for the solution of a "forward" equation of the Hamilton-Jacobi-Bellman-Isaacs type or as a solution to an evolution equation of the integral funnel type. Ellipsoidal-valued approximations of the reach sets are introduced.

Introduction

In this report we discuss the computation of domains reachable by a controlled process through available controls. This is one of the key issues in control theory which is usually studied assuming the system model is known. However, in reality the system may be subject to unknown disturbances and the information on its parameters may not be complete. It then becomes necessary to describe the states reachable by the system *despite the disturbances* or incomplete information or, if exact reachability is impossible, to find the guaranteed errors for reachability. A crucial element in obtaining satisfactory reachability properties is that the system has to be governed by *closed – loop controls.* This gives substantial improvement as compared to nonanticipative ("minmax") open-loop controls. Here the reachable sets are described either as level sets for the solution to a first order PDE of the "forward" HJBI (Hamilton-Jacobi-Bellman-Isaacs) type, or as a set-valued solution to an evolution equation of the integral funnel type (see [2], [4], [11], [5], [7]). External and internal ellipsoidal-valued approximations of reach sets are then described using the relations of [6], [8].

1 The Reachability Problem

Consider the system

$$\dot{x} = A(t)x + B(t)u + C(t)v(t), \tag{1}$$

---

*Research supported by National Science Foundation Grant ECS 9725148

with continuous matrix coefficients $A(t), B(t), C(t)$. Here $x \in \mathbb{R}^n$ is the *state* and $u \in \mathbb{R}^p$ is the *control* that may be selected either as an *open loop control* OLC – a Lebesgue-measurable function $u(t)$ of time t, or as a *closed-loop control* CLC – a *set-valued strategy* $\mathcal{U}(t,x)$, restricted respectively by the inclusions

$$u(t) \in \mathcal{P}(t), \; u = \mathcal{U}(t,x) \subseteq \mathcal{P}(t) \; a.e., \tag{2}$$

Symbol $v \in \mathbb{R}^q$ stands for the unknown *input disturbance* with values

$$v(t) \in \mathcal{Q}(t).$$

$\mathcal{P}(t), \mathcal{Q}(t)$ are set-valued functions with convex compact values, continuous in time.

The class of OLC's $u(\cdot)$ bounded by the first inclusion of (2) is denoted by U_O and the class of bounded input disturbances $v(\cdot)$ as V_O. The strategies $\mathcal{U}$ are taken to be in U_C – the class U_C of CLC's that consists of multivalued maps $\mathcal{U}(t,x)$ bounded by the second inclusion (2), and such that ensure the existence and extendability of solutions to equation (1), $u = \mathcal{U}(t,x)$, (equation (1) then turns into a differential inclusion - DI) for any Lebesgue-measurable function $v(\cdot)$. Such, for example, is the class of set-valued functions with values in *comp*$\mathbb{R}^n$, upper semicontinuous in x and continuous in t.

Under the given conditions equation (1) is said to describe an *uncertain system* of the *linear-convex type*.

Definition 1.1 *The reach set $\mathcal{X}(\tau, t_0, \mathcal{X}^0; \mu)$ of system (1), at time τ, under uncertain input disturbances $v(t)$, is the set of all points x, for each of which there exists a CLC strategy $\mathcal{U} \in U_C$, such that for every $v(\cdot) \in V_O$ there exists a solution tube $\mathcal{X}(t, t_0, x^0 | \mathcal{U}, v(\cdot))$, $x^0 \in \mathcal{X}^0$, to the differential inclusion*

$$\dot{x} \in A(t)x + B(t)\mathcal{U}(t,x) + C(t)v(t), \;\; x(t_0) = x^0, \; t \geq t_0 \tag{3}$$

whose cross-section ("cut") $\mathcal{X}(\tau, t_0, x^0 | \mathcal{U}, v(\cdot))$ at time $t = \tau$ satisfies the relation

$$x \in \mathcal{X}(\tau, t_0, x^0 | \mathcal{U}, v(\cdot)) + \mu \mathcal{B}(0).$$

Problem I (The reachability problem), is to find (calculate) the set $\mathcal{X}(\tau, t_0, \mathcal{X}^0; 0)$ or, if this set is empty, to find the set $\mathcal{X}(\tau, t_0, \mathcal{X}^0; \mu)$ for some $\mu > 0$.

Note that strategy $\mathcal{U}$ of (2) is *one and the same for all* $v(\cdot) \in V_O$. Clearly, for μ sufficiently large, the last set will always be nonempty. It therefore makes sense to look for the smallest of such μ. We shall reduce the reachability problem to some optimization problems.

2 Reachability through Dynamic Optimization

Problem 2.1 Find the *value function*

$$\mathcal{V}(\tau, x) \; = \; \min_{\mathcal{U}} \max_{x(\cdot)} \{\phi(x(t_0)) | x(\tau) = x, \; \mathcal{U} \in U_C, \; x(\cdot) \in \mathcal{X}_{\mathcal{U}}(\cdot)\}, \tag{4}$$

where $\phi(x) = d^2(x, \mathcal{X}^0) = \mathcal{V}(t_0, x)$, $\;d(x, \mathcal{X}^0) = \min\{(x - z, x - z)^{1/2} | z \in X^0\}$ and $\mathcal{X}_{\mathcal{U}}(\cdot) = \mathcal{X}(\cdot, \tau, x | \mathcal{U})$ is the tube of all trajectories $x(\cdot)$ of the differential inclusion

$$\dot{x} \in A(t)x + B(t)\mathcal{U}(t,x) + C(t)\mathcal{Q}(t), \;\; x(\tau) = x, \; t \in [t_0, \tau],$$

generated by the strategy $\mathcal{U} \in U_C$.

The value function $\mathcal{V}(\tau, x) = \mathcal{V}(\tau, x|\mathcal{V}(t_0, \cdot))$ satisfies *The Principle of Optimality under Uncertainty* (see [7]) which is formulated through a semi-group property for the respective mapping:

$$\mathcal{V}(\tau, x|\mathcal{V}(t_0, \cdot)) = \mathcal{V}(\tau, x|\mathcal{V}(t, \cdot \mid \mathcal{V}(t_0, \cdot))),\ t_0 \leq t \leq \tau. \tag{5}$$

The last relation leads to a formal *forward-type* HJBI equation for $\mathcal{V}(t, x)$, which is

$$\partial\mathcal{V}/\partial t \ + \ \Big(\partial\mathcal{V}/\partial x, A(t)x\Big) \ + \ \rho\Big(\partial\mathcal{V}/\partial x|B(t)\mathcal{P}(t)\Big) - \rho\Big(\partial\mathcal{V}/\partial x|C(t)\mathcal{Q}(t)\Big) = 0, \tag{6}$$

with boundary condition $\mathcal{V}(t_0, x) \ = \ \phi(x) = d^2(x, \mathcal{X}^0)$. Here

$$\rho(l|\mathcal{Q}) = \max\{(l, q)|q \in \mathcal{Q}\}$$

stands for the *support function* of set $\mathcal{Q}$, [10].

Theorem 2.1 *Function $\mathcal{V}(t, x)$ is a generalized ("viscosity") solution of (6).*

Such solutions are explained, for example, in [3]. The principal result concerning $\mathcal{X}(\tau, t_0, \mathcal{X}^0; \mu)$ is as follows.

Theorem 2.2 *The reach set $\mathcal{X}(\tau, t_0, \mathcal{X}^0; \mu)$ (under closed-loop control) is the level set*

$$\mathcal{X}(\tau, t_0, \mathcal{X}^0; \mu) = \{x : \mathcal{V}(\tau, x) \leq \mu\}.$$

This theorem was proved in [7]. However, the calculation of set $\mathcal{X}[\tau] = \mathcal{X}(\tau, t_0, \mathcal{X}^0; \mu)$ and its evolution in τ is not simple, especially, since $\mathcal{X}[\tau]$ is set-valued. In principle it could be done either by finding $\mathcal{X}[\tau]$ through $\mathcal{V}(t, x)$ – the solution of the PDE (6)– or by finding it directly from a "funnel"- type evolution equation with set-valued solutions, as given below. To overcome the difficulties we shall later indicate some external and internal ellipsoidal approximations of $\mathcal{X}[\tau]$ based on the relations given in [6], [8].

An important feature is that $\mathcal{V}(\tau, x)$ may be obtained as a limit with $k \to \infty$ of a superposition $V_k^+(\tau, x|\mathcal{V}(t_0, \cdot))$ of the solutions of k open-loop minmax control problems. These are defined as follows.

3 Reachability under Open-loop Controls with Corrections

Taking the interval, $T = [t_0, \tau]$, introduce a partition

$$\Sigma_k = \{t_0 = \tau_0, \tau_1, \ldots, \tau_k, \tau = \tau_{k+1}\},\ \tau_i - \tau_{i-1} = \sigma_i \geq 0,\ i = 1, \ldots, k+1,$$

that divides T into $k + 1$ intervals

$$T_0 = [t_0, \tau_1),\ T_2 = [\tau_1, \tau_2), \ldots, T_k = [t_1 - \tau_k, \tau],$$

where

$$\tau_i = t_0 + \sum_{j=1}^{i} \sigma_j,\ \ i = 1, \ldots, k,$$

will serve as *the points of corrections* of the open-loop controls.

Problem 2- k. Solve the following consecutive optimization problems.

Assuming $V_k^+(t_0, x) = d^2(x, X^0)$, find

$$V_k^+(\tau_1, x|V_k^+(t_0, \cdot)) = \min_u \max_v \{V_k^+(t_0, x(t_0))\ |x(\tau_1) = x,\ u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_1\},$$

then consecutively

$$V_k^+(\tau_2, x|V_k(t_1, \cdot)) = \min_u \max_v \{V_k(t_1, x(t_1))|x(\tau_1) = x,\ u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_2\},$$

then, further, for $i = 3, \ldots, k$, find

$$V_k^+(\tau_i, x|V_{i-1}^+(\tau_{i-1}, \cdot)) =$$

$$= \min_u \max_v \{V_k^+(\tau_{i-1}, x(\tau_{i-1})|x(\tau_i) = x,\ u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_i\},$$

and finally

$$V_k^+(\tau, x|V_k^+(\tau_k, \cdot)) =$$

$$\min \max_u \{V_k^+(\tau_k, x(\tau_k))|x(\tau) = x,\ u(t) \in \mathcal{P}(t), v(t) \in \mathcal{Q}(t), t \in T_{k+1}\},$$

so that

$$V_k^+(\tau, x|V_k^+(t_0, \cdot)) = \tag{7}$$

$$V_k^+(\tau, x|\ V_k^+(\tau_k, x(\cdot)| \ldots |V_k^+(\tau_1, x(\cdot)|V_k^+(t_0, \cdot))\ldots)$$

Here it is presumed that at each point τ_i it is possible to measure the exact value $x(\tau_i)$, while on each interval $T_{i+1} = [\tau_i, \tau_{i+1})$ the open-loop control $u(t)$ is selected *one an the same for all* $v(t)$, as a *nonanticipative control.*

A similar function

$$V_k^-(\tau, x|\ V_k^-(\tau_k, x(\cdot)| \ldots |\tau, V_k^-(\tau_1, x(\cdot)|V_k^-(t_0, \cdot))\ldots)$$

is defined through a substition of all operations min max by max min. Here it is again presumed that at each point τ_i it is possible to measure the exact value $x(\tau_i)$, but on each interval $T_{i+1} = [\tau_i, \tau_{i+1})$ the open-loop control $u(t)$ is selected *depending on* $v(t)$, as an *anticipative control.*

The following convergence theorem is true.

Theorem 3.1 *Suppose functions* $V_k^+(\tau, x)$ *are proper convex ([10]). Consider a sequence of partitions* Σ_k *with* $k \to \infty$ *and*

$$\max\{\sigma_i\ :\ i = 1, \ldots, k+1\} \to 0, \sum_{i=1}^{k+1} \sigma_i = \tau - t_0. \tag{8}$$

(Without loss of generality the sequence Σ_k *may be always considered monotone, namely, achieved by adding new points of correction to the old ones).*

Then there exists a pointwise limit

$$\lim_{k \to \infty} V_k^+(\tau, x) = \mathcal{V}_0^+(\tau, x). \tag{9}$$

This limit does not depend on the type of partititons Σ_k.

A similar convergence theorem is also true for proper convex functions $V_k^-(\tau, x)$, namely, under conditions (8),

$$\lim_{k\to\infty} V_k^-(\tau, x) = \mathcal{V}_0^-(\tau, x).$$

Under the conditions of the above the important fact lies in in the following, [7]. [1]

Theorem 3.2 *The limits* $\mathcal{V}_0^+(\tau, x), \mathcal{V}_0^-(\tau, x)$ *coincide with* $\mathcal{V}(\tau, x)$, *namely*

$$\mathcal{V}_0^-(\tau, x) = \mathcal{V}(\tau, x) = \mathcal{V}_0^+(\tau, x).$$

Lemma 3.1 *The following inequalities are true*

$$V_0^-(\tau, x) \leq \ldots \leq V_k^-(\tau, x) \leq \ldots \leq \mathcal{V}(\tau, x) \leq \ldots \leq V_k^+(\tau, x) \leq \ldots \leq V_0^+(\tau, x). \qquad (10)$$

Introducing the level sets

$$\mathcal{X}_k^+(\tau, t_0, \mu) = \{x : V_k^+(t, x) \leq \mu\}, \ \mathcal{X}_k^-(\tau, t_0, \mu) = \{x : V_k^-(t, x) \leq \mu\},$$

one may observe that (10) yields the next relations.

Theorem 3.3 *The following inclusions are true.*

$$\begin{aligned} \mathcal{X}_0^+(\tau, t_0, \mu) \subseteq \ldots \subseteq \mathcal{X}_k^+(\tau, t_0, \mu) \subseteq \ldots \subseteq \mathcal{X}(\tau, t_0, \mu) & \qquad (11) \\ \subseteq \ldots \subseteq \mathcal{X}_k^-(\tau, t_0, \mu) \subseteq \ldots \mathcal{X}_0^+(\tau, t_0, \mu). & \end{aligned}$$

Here $\mathcal{X}_k^+(\tau, t_0, \mu)$, $\mathcal{X}_k^-(\tau, t_0, \mu)$ may be interpreted as *reach sets under open-loop controls with k corrections* at instants τ_i, while set $\mathcal{X}(\tau, t_0, \mu)$, as indicated above, is the *reach set under closed loop control.* All these sets are convex and compact.

As a consequence of (11), we come to the next result

$$\lim_{k\to\infty} h(\mathcal{X}_k^+(\tau, t_0, \mu), \mathcal{X}(\tau, t_0, \mu)) = 0, \ \lim_{k\to\infty} h(\mathcal{X}_k^-(\tau, t_0, \mu), \mathcal{X}(\tau, t_0, \mu)) = 0, \qquad (12)$$

where

$$h(\mathcal{Q}, \mathcal{P}) = \max\{h_+(\mathcal{Q}, \mathcal{P}), h_+(\mathcal{P}, \mathcal{Q})\}, h_+(\mathcal{Q}, \mathcal{P}) = \min\{\epsilon > 0 : \mathcal{Q} \subseteq \mathcal{P} + \epsilon\mathcal{B}(0)\}$$

stand for the Hausdorff distance h and semidistance h_+. The reach set $\mathcal{X}[\tau] = \mathcal{X}(\tau, t_0, \mu)$ may thus be calculated as the Hausdorff limit of open-loop reach sets under corrections, when the number of these tends to infinity due to a monotone sequence of partitions Σ_k.

Remark 3. There is one particular case when $V_0^-(\tau, x) \equiv \mathcal{V}(\tau, x)$ and therefore all the functions $V_k^-(\tau, x)$ of the inequalities (10) coincide. This is when system (1) satisfies the so-called *matching conditions*:

$$B(t)\mathcal{P}(t) \equiv \alpha(t)C(t)\mathcal{Q}(t), \ \alpha(t) \in [0, 1).$$

The results of this section are not the only approach to the reachability problem. Namely, the set $\mathcal{X}[\tau]$ may also be defined as the solution to a certain type of evolution "funnel equation", [1], [6].

---

[1] In this case there also holds the property of *epiconvergence* of $V_k^+(\tau, x)$ and $V_k^-(\tau, x)$ to $\mathcal{V}_0(\tau, x)$ (see [10]).

4 Reachability through Funnel Equations

Consider the evolution equations

$$\lim_{\sigma\to 0}\sigma^{-1}h_+(\mathcal{X}^+(t+\sigma),((I+\sigma A(t))\mathcal{X}^+(t)\dot{-}\sigma(-C(t)\mathcal{Q}(t)))+\sigma B(t)\mathcal{P}(t))=0, \qquad (13)$$

with initial condition

$$\mathcal{X}^+(t_0)=\mathcal{X}^0+\sigma K\mathcal{B}(0),$$

and

$$\lim_{\sigma\to 0}\sigma^{-1}h_+(\mathcal{X}^-(t+\sigma),((I+\sigma A(t))\mathcal{X}^-(t)+\sigma B(t)\mathcal{P}(t))\dot{-}(-C(t)\mathcal{Q}(t)))=0, \qquad (14)$$

with

$$\mathcal{X}^-(t_0)=\mathcal{X}^0,$$

Here $\mathcal{P}_1\dot{-}\mathcal{P}_2=\{p:\mathcal{P}_2+p\subseteq\mathcal{P}_1\}$ is the geometric ("Minkowski") difference of sets $\mathcal{P}_1$ and $\mathcal{P}_2$. Under some *regularity assumptions* indicated below each of these equations has a unique set-valued solution ($\mathcal{X}^+(t)$ and $\mathcal{X}^-(t)$ accordingly), which is generated by the indicated initial condition and maximal with respect to inclusion, satisfying the respective equation almost everywhere (see similar situations in [5], [6]). [2]

The *regularity assumptions* mentioned in the previous lines are as follows.

Assumption 4.1 *There exist continuous vector functions $\beta^+(t),\beta^-(t)\in\mathbb{R}^n$, and numbers $\epsilon>0,\sigma_0>0$ such that for any $\sigma\le\sigma_0$ the following inclusions would be true:*

$$\beta_1(t)+\epsilon\mathcal{B}(0)\subseteq\mathcal{X}_\sigma^+(t),\quad \beta_1(t)+\epsilon\mathcal{B}(0)\subseteq\mathcal{X}_\sigma^-(t),\quad t_0\le t\le\tau,$$

where $\mathcal{X}_\sigma^+(t),\mathcal{X}_\sigma^-(t)$ are the solutions to the recurrence equations

$$\mathcal{X}_\sigma^+(t+\sigma)=((I+\sigma A(t))\mathcal{X}_\sigma^+(t)\dot{-}(-C(t)\mathcal{Q}(t)))+\sigma B(t)\mathcal{P}(t),\ \mathcal{X}_\sigma^+(t_0)=\mathcal{X}^0+K\sigma\mathcal{B}(0), \qquad (15)$$

and

$$\mathcal{X}_\sigma^-(t+\sigma)=(I+\sigma A(t))\mathcal{X}_\sigma^-(t)+\sigma(B(t)\mathcal{P}(t)\dot{-}(-C(t)\mathcal{Q}(t))),\ \mathcal{X}_\sigma^-(t_0)=\mathcal{X}^0, \qquad (16)$$

Theorem 4.1 *Under Assumption 4.1 the following equalities are true*

$$\lim_{\sigma\to 0}h(\mathcal{X}^+(\tau),\mathcal{X}_\sigma^+(\tau))=0,\ \lim_{\sigma\to 0}h(\mathcal{X}^-(\tau),\mathcal{X}_\sigma^-(\tau))=0,$$

and

$$\mathcal{X}[\tau]=\mathcal{X}^+(\tau)=\mathcal{X}^-(\tau).$$

This means equations (13), (14) describe the evolution of set $\mathcal{X}[\tau]=\mathcal{X}(\tau,t_0,X^0)$ in time, while the recurrence equations (15), (16)allow to approximate the exact solutions. The role of sets $\mathcal{X}_\sigma^+(\tau),\mathcal{X}_\sigma^-(\tau)$ is similar to that of the alternated integral sums of the first and second kind in [5], [7]. Despite the available formal solution schemes, the real calculations of sets $\mathcal{X}(\tau,x)$ along these schemes are difficult even in the absence of uncertainty [9]. We therefore indicate some ellipsoidal representations which allow effective approximations of these sets.

---

[2] In order that equation (13) would be defined properly, one must require that constant $K>0$ is chosen such that for some σ_0 we have $((I+\sigma A(t_0))\mathcal{X}^0+\sigma K\mathcal{B}(0))\dot{-}(\sigma C(t_0)\mathcal{Q}(t_0)+\sigma\mathcal{B}(0))\neq\emptyset,\forall\sigma\le\sigma_0$.

5 Ellipsoidal Representations of Reach Sets

We shall construct the ellipsoidal representations following the scheme of [6], Section 3.5, but with the formulas of paper [8]. In this Section we assume $\mathcal{P}(t)$, $\mathcal{Q}(t)$ and $\mathcal{X}^0$ to be ellipsoidal-valued:

$$\mathcal{P}(t) = \mathcal{E}(p(t), P(t)),\ \mathcal{Q}(t) = \mathcal{E}(q(t), Q(t)),\ \mathcal{X}^0 = \mathcal{E}(x^0, X^0),$$

where $\mathcal{E}(x, X)$ is an ellipsoid with center x and configuration (shape) matrix $X = X' > 0$, so that its support function $\rho(l|\mathcal{E}(x, X)) = (l, x) + (l, Xl)^{1/2}$. Our present assumption is justified, since any convex compact set may be presented as an intersection of ellipsoids.

Let us look for ellipsoidal-valued solutions to the funnel equation (14). Rewriting it in new notations, we have

$$\lim_{\sigma\to 0} \sigma h_+(\mathcal{X}^-(t+\sigma) - \sigma C(t)\mathcal{E}(q(t), Q(t)), (I + \sigma A(t))\mathcal{X}^-(t) + \sigma B(t)\mathcal{E}(p(t), P(t))) = 0,$$
$$\mathcal{X}^-(t_0) = \mathcal{E}(x^0, X^0).$$

The existence of an ellipsoidal solution to the last equation follows from the relation

$$\mathcal{X}^-(t+\sigma) - \sigma C(t)\mathcal{E}(q(t), Q(t))$$
$$\subseteq (I + \sigma A(t))\mathcal{X}^-(t) + \mathcal{E}(\sigma B(t)p(t), \sigma^2 B(t)P(t)B'(t)) + o(\sigma)\mathcal{B}(0), \tag{17}$$

which follows from the definition of h_+, and further on, for the right-hand side of (17),

$$(I + \sigma A(t))\mathcal{X}^-(t) + \sigma\mathcal{E}(B(t)p(t), B(t)P(t)B'(t)) \subseteq$$
$$\subseteq (I + \sigma A(t))(\mathcal{E}(x(t), X^-(t)) + \sigma\mathcal{E}(B(t)p(t), B(t)P(t)B'(t)) \subseteq$$
$$\subseteq \mathcal{E}(z_+(t), Z_+(t)) + o_1(\sigma)\mathcal{B}(0), \tag{18}$$

where

$$z_+(t) = (I + \sigma A(t))x(t) + \sigma B(t)p(t),$$
$$Z_+(t) = X^-(t) + \sigma(A(t)X^-(t) + X^-(t)A'(t)) + \sigma(\pi(t))^{-1}B(t)P(t)B'(t)$$

and $p(t) > 0$ is any positive function. Here and further $o(\sigma)/\sigma \to 0$, $o_i(\sigma)/\sigma \to 0$ with $\sigma \to 0$.

On the other part, for the left-hand side of (17), we have

$$\mathcal{X}^-(t+\sigma) - \sigma C(t)\mathcal{E}(q(t), Q(t)) \supseteq \mathcal{E}(x(t+\sigma), X^-(t+\sigma)) + \mathcal{E}(-\sigma C(t)q(t), \sigma^2 C(t)Q(t)C'(t)) \tag{19}$$
$$\supseteq \mathcal{E}(z_-(t), Z_-(t)) + o_2(t)\mathcal{B}(0),$$

where

$$z_-(t) = x(t+\sigma) - \sigma C(t)q(t),$$
$$Z_-(t) = (X_*^-(t+\sigma) + \sigma S(t)(C(t)Q(t)C'(t))^{1/2})'(X_*^-(t+\sigma) + \sigma S(t)(C(t)Q(t)C'(t))^{1/2}),$$

with $X_*^{-\prime}(t)X_*^-(t) \equiv X^-(t)$, and $S(t)$ being any orthogonal matrix-valued function.

The desired ellipsoidal-valued solution to (18) is obtained by equalizing $z_-(t) \equiv z_+(t)$, $Z_-(t) \equiv Z_+(t)$, dividing both relations by $\sigma > 0$ and passing to the limit with $\sigma \to 0$. This gives

$$\dot{x}(t) = A(t)x(t) + B(t)p(t) + C(t)q(t), \tag{20}$$

$$\dot{X^-} = A(t)X^- + X^-A'(t) + \pi(t)X^-(\pi(t))^{-1}B(t)P(t)B'(t) - \tag{21}$$
$$-X_*^{-'}S(t)(C(t)Q(t)C'(t))^{1/2} - ((C(t)Q(t)C'(t))^{1/2}S'(t)X_*^{-'}$$

with

$$x(t_0) = x^0, \;\; X^-(t_0) = X^0. \tag{22}$$

Similarly, following equation (14), we may introduce an ellipsoidal-valued function $\mathcal{E}(x(t), X^+(t)) \subseteq \mathcal{X}[t]$, where $x(t)$ is the same as in (20) and

$$\dot{X^+} = A(t)X^+ + X^+A'(t) - \pi_-(t)X^+ - (\pi_-(t))^{-1}C(t)Q(t)C'(t) \tag{23}$$
$$+X_*^{+'}S(t)(B(t)P(t)B'(t))^{1/2} + ((B(t)P(t)P'(t))^{1/2}S'(t)X_*^+$$

with initial conditions $X^0 = X^+(t_0)$ and with $X_*^{+'}(t)X_*^+(t) \equiv X^+[t]$.

Theorem 5.1 *The ellipsoids $\mathcal{E}(x(t), X^+(t)), \mathcal{E}(x(t), X^-(t))$ ensure the inclusions*

$$\mathcal{E}(x(t), X^+(t)) \subseteq \mathcal{X}[t] \subseteq \mathcal{E}(x(t), X^-(t)), \tag{24}$$

whatever be the functions $\pi(t) > 0$ and the orthogonal matrix-valued $S(t)$. The function $x(t) \in \mathbb{R}^n$ satisfies (20), and $X^-(t), X^+(t) \in \mathbb{R}^{n\times n}$ satisfy the equations (21), (23), with initial conditions $X^-(t_0) = X^0 = X^+(t_0)$.

An important property is true for these ellipsoidal approximations under the regularity assumptions of the above.

Theorem 5.2

$$\cup\{\mathcal{E}(x(\tau), X^+(\tau))|\pi(\cdot), S(\cdot)\} = \mathcal{X}[\tau] = \cap\{\mathcal{E}(x(\tau), X^-(\tau))|\pi(\cdot), S(\cdot)\} \tag{25}$$

over all possible functions $p(t) > 0$ and orthogonal matrix-valued $S(t)$.

This means there exist *exact ellipsoidal representations* for the closed-loop reach sets.

A further property similar to the one introduced for systems without uncertainty in [8] is true. Namely, consider the curve $l(t) = s(t, t_0, l), l \in \mathbb{R}^n$, where $s(t, t_0, l)$ is the solution to the "adjoint system" $ds/dt = -A'(t)s, \;\; s(t_0) = l$ and select the parameters $p(t), S(t)$ in (21) due to relations

$$\pi(t) = (l(t), P(t)l(t))^{1/2}/(l(t), X^-(t)l(t))^{1/2}$$

$$S(t)(C(t)Q(t)C'(t))^{1/2}l(t) = ((l(t), C(t)Q(t)C'(t)l(t))/(l(t), X^-(t)l))^{1/2}X_*^-(t)l(t).$$

Then the external ellipsoidal approximations $\mathcal{E}(x(t), X^-(t)$ will be *tight*, namely, the equality

$$\rho(l(t)|\mathcal{X}[t]) = (l(t), x(t)) + (l(t), X^-(t)l(t))^{1/2} = \rho(l(t)|\mathcal{E}(x(t), X^-(t))),$$

will be true for all $t \in [t_0, \tau]$ and every $l \in \mathbb{R}^n$ that generates the curve $l(t) = s(t, t_0, l)$. The tube $\mathcal{X}[t]$ will lie entirely within a family of finite dimensional ellipsoidal-valued tubes each of which externally touches the surface of $\mathcal{X}[t]$ along a continuous curve $x_*(t) = x(t) + X^-(t)l(t)((l(t), X^-(t)l(t))^{-1/2}$.

A similar property is also true for internal approximations.

Conclusion

In this paper we introduce the notion of reach sets under feedback (closed – loop) control and set-membership input uncertainty, with hard bounds on the unknowns. We indicate that they may be obtained as level sets for the viscosity solutions to a *forward* HJBI (Hamilton-Jacobi-Bellman-Isaacs) equation or as set-valued solutions to certain evolution equations of the "funnel" type. These sets naturally turn to be larger than those for a similar nonanticipative minmax problem of open–loop control. Their calculation could be achieved by using the techniques of ellipsoidal calculus, for example, by embedding the desired reachability tube $\mathcal{X}[t]$ into a family of tight external ellipsoidal tubes that touch $\mathcal{X}[t]$ along some continuous curves which fully cover its surface. The demand for calculating reach sets under disturbances and uncertainty is motivated by applied problems of control for new types of complex transition systems.

References

[1] AUBIN J-P., FRANKOWSKA H., *Set-valued Analysis*, Birkhäuser, Boston, 1991.

[2] BASAR T., BERNHARD P., *H^∞ Optimal Control and Related Minimax Design Problems*, 2-nd. ed., Birkhäuser, Boston, 1997.

[3] FLEMING W.H., SONER H.M.,*Controlled Markov Processes and Viscosity Solutions*, Springer - Verlag, 1993.

[4] KRASOVSKI N.N., SUBBOTIN A.N., *Positional Differential Games*, Springer-Verlag, 1988.

[5] KURZHANSKI A.B., "Pontryagin's alternated integral in the theory of control synthesis". Proc. Steklov Math. Inst., v.224, 1999, (Engl. Transl.), pp. 234 - 248.

[6] KURZHANSKI A.B., VALYI I., *Ellipsoidal Calculus for Dynamics and Control*, Birkhäuser, 1998.

[7] KURZHANSKI A.B., VARAIYA P., "On reachability under uncertainty." SIAM Journal on Control and Optimization (to appear).

[8] KURZHANSKI A.B., VARAIYA P. "Ellipsoidal techniques for reachability analysis". Lecture Notes in Computer Sciences, 1790, Hybrid systems: computation and control, N.Lynch, B.L.Krogh eds.,Springer, 1999, pp.202-214.

[9] PURI A., BORKAR V. and VARAIYA P., "ϵ - Approximations of Differential Inclusions" Lecture Notes in Computer Sciences, 1066, Hybrid systems, R.Alur, T.A.Henzinger, E.D.Sontag eds., Springer, 1996, pp.362-376.

[10] ROCKAFELLAR R.T., WETS R.J.B., *Variational Analysis*, Springer-Verlag, 1998.

[11] VARAIYA P., LIN J., "Existence of Saddle Points in Differential Games", SIAM Journal on Control and Optimization, v.7,1, 1969, pp.142 - 157.

On Optimal Input Design for Parameter Estimation Problems in PDE's*

P.-L. Chow
Department of Mathematics,
Wayne State University
Detroit MI, 48202 USA
plchow@math.wayne.edu;

R. Khasminskii
Department of Mathematics
Wayne State University
Detroit MI, 48202 USA
rafail@math.wayne.edu

Abstract

We consider the problem of optimal input design for the asymptotically efficient estimation of a parameter for some parabolic PDE's. Asymptotically efficient estimators are also proposed.

1 Introduction

The optimal choice of an input signal in the parameter identification problems for linear systems is important in many applications. For systems described by Ordinary Differential Equations (ODE's) this problem was considered , in particular, in [1,8-14]. Recently the approach of afore-mentioned papers was generalized in [15], where the asymptotically optimal input signal was proposed for the stable, neutral and unstable systems described by ODE's of which solutions are observed in the White Gaussian Noise (WGN).

In this paper we consider an analogous problem for the systems described by Partial Differential equations (PDE's).At the beginning, in section 2, we consider a rather general approach to the optimal input design problem for the systems, described by PDE's and more general problem under the assumption that the solution of the system is observed in a small WGN.

In the next two sections we consider examples of the evolutional systems assuming that the solutions of these systems are observed in the moderate WGN with a long observation interval. We find that the situation is very different for parabolic and hyperbolic differential equations. For special examples we not only find the lower bound for the estimation performance, but also construct asymptotically optimal input signal and asymptotically efficient estimators.

---

*This research is supported in part by NSF Grant No. DMS-9971608

2 Optimal Input Signal for the Small Observation Noise.

Let $L(x,\theta)$, $x \in D \subset \mathbf{R}^n$ be a linear operator with $\theta \in \Theta \subset \mathbf{R}^1$ (differential operator, for instance), such that the equation

$$L(x,\theta)S = u(x) \tag{1}$$

has a unique solution subject to the initial/boundary conditions $S|_{\partial D} = 0$ for any u with $||u|| \le 1$ (here and below we denote by $||.||$ the L_2-norm in D). This solution can be written as

$$S = G(\theta)u(x), \tag{2}$$

where $G(\theta)$ is a linear operator.We assume that θ is unknown parameter, the operator $L(x,\theta)$ and the solution of (1) are differentiable in θ for all $||u|| \le 1$ (u does not depend on θ). Assume also that the observed random measure $Y_\varepsilon(\Gamma), \Gamma \in D$, has the form

$$Y_\varepsilon(dx) = KS(x,\theta)dx + \varepsilon W(dx) \tag{3}$$

As usual (see, e.g. [7], [4]) we interpret (3) in the generalized sense: the observable elements are random linear functionals

$$\int_D f(x)Y_\varepsilon(dx) = \int_D f(x)KS(x,\theta)dx + \varepsilon \int_D f(x)W(dx). \tag{4}$$

Here $W(.)$ is a Gaussian random measure on D such that for any Borel sets Γ_i

$$\mathbf{E}W(\Gamma_1) = 0; \mathbf{E}W(\Gamma_1)W(\Gamma_2)) = \Lambda(\Gamma_1 \bigcap \Gamma_2),$$

where $\Lambda(.)$ is a Lebesgue measure in $\mathbf{R}^n$, K is some bounded linear operator $L_2(D) \mapsto L_2(D)$, $f \in L_2(D)$. The problem is: based on the observation $Y_\varepsilon(.)$ to find the best input u for the parameter estimation with the constraint on its L_2 -norm:

$$||u|| \le 1. \tag{5}$$

Denote by $\mathcal{U}$ the set of inputs that do not depend on θ and satisfy the restriction (5). We assume that the following condition is valid.

Assumption **A**. *For all $\theta \in \Theta$ and $u \in \mathcal{U}$, the function $KS(.,\theta)$ is differentiable with respect to θ in $L_2(D)$ and this derivative is equal to $K\frac{\partial}{\partial\theta}S(.,\theta)$.*

Denote by $\mathbf{P}_{u,\theta}^{(\varepsilon)}(.)$ the probability measure generated by the observation (3). It is well known that under assumption **A**, for different θ, the measures $\mathbf{P}_{u,\theta}^{(\varepsilon)}(.)$ are absolutely continuous (see, e.g.[3], section 7.4), and

$$\begin{aligned} \log \frac{d\mathbf{P}_{u,\theta+\varepsilon h}^{(\varepsilon)}}{d\mathbf{P}_{u,\theta}^{(\varepsilon)}}(Y_\varepsilon) &= -\frac{1}{2\varepsilon^2}\int_D |K(S(x,\theta+\varepsilon h) - S(x,\theta))|^2 dx \\ &+\frac{1}{\varepsilon}\int_D K(S(x,\theta+\varepsilon h) - S(x,\theta))W(dx) \end{aligned} \tag{6}$$

Letting $\varepsilon \to 0$ and using assumption **A** again, we have the equation (we denote below by $(.,.)$ the inner product in $L_2(D)$)

$$\log \frac{d\mathbf{P}^{(\varepsilon)}_{u,\theta+\epsilon h}}{d\mathbf{P}^{(\varepsilon)}_{\theta}}(Y_\varepsilon) = -\frac{1}{2}h^2(K^*KS'_\theta(.,\theta), S'_\theta(.,\theta)) + h\int_D KS'_\theta(x,\theta)W(dx) + o(1). \tag{7}$$

(Here and below we denote by B^* the adjoint operator to B). So the family of experiments generated by (3)has the LAN (Local Asymptotic Normality) property (see, e.g., [6]) with the Fisher information

$$I(\theta,u) = (K^*KS'_\theta(.,\theta), S'_\theta(.,\theta)) \tag{8}$$

Further, taking (2) into account, we can rewrite (5) in the form

$$I(\theta,u) = (K^*KG'_\theta(.,\theta)u, G'_\theta(.,\theta)u) \tag{9}$$

It is difficult to find the explicit form of $G'_\theta(x,\theta)$ even for the case when $L(x,\theta)$is an ordinary differential operator ($n=1$). But the equation (9) can be written in a more convenient form. For this purpose let us differentiate (1). Then we obtain the equation

$$L(x,\theta)S'_\theta(x,\theta) + L'_\theta(x,\theta)S(x,\theta) = 0 \tag{10}$$

and the boundary/initial condition $S'_\theta(x,\theta)|_{\partial D} = 0$. Thus (see(2)) the function $S'_\theta(x,\theta)$ can be represented as

$$S'_\theta(x,\theta) = -G(\theta)L'_\theta(x,\theta)S(x,\theta),$$

and

$$\begin{aligned} I(\theta,u) &= (K^*KG(\theta)L'_\theta(.,\theta)G(\theta)u, G(\theta)L'_\theta(.,\theta)G(\theta)u) \\ &= (G^*(\theta)L'^*_\theta(.,\theta)G^*(\theta)K^*KG(\theta)L'_\theta(.,\theta)G(\theta)u, u) = (Au,u). \end{aligned} \tag{11}$$

Assumption **B.** *The operator $A(\theta)$: $L_2(D) \longmapsto L_2(D)$ is bounded for all $\theta \in \Theta$ and*

$$\sup_{u\in\mathcal{U}}(A(\theta)u,u) = \lambda_0(\theta) = (A(\theta)u_0,u_0).$$

Thus $\lambda_0(\theta)$ is the maximal eigenvalue for the operator $A(\theta)$ and $u_0(\theta)$ is the corresponding eigenfunction. It is well known (see, [6], subsection 2.12) that for the LAN families of experiments the asymptotically minimax (for $\varepsilon \to 0$) lower bound for the estimation risks is less for the families with greater Fisher information. From this fact and (11) we obtain the theorem generalizing results of [15], where only the case of ODE operators L was considered for the quadratic loss function.

Theorem 1 *Let $l(x)$ be any symmetric and increasing, for $x > 0$, function $R \longmapsto R^+$. Then for any estimator $\hat{\theta}_\varepsilon$ of the parameter θ, any $\theta_0 \in \Theta$ and any $\delta > 0$,*

$$\liminf_{\varepsilon\to 0}\inf_{u\in\mathcal{U}}\sup_{\{\theta:|\theta-\theta_0|<\delta\}}\mathbf{E}_{\theta,u}l(\frac{1}{\varepsilon}(\hat{\theta}_\varepsilon-\theta)) = \liminf_{\varepsilon\to 0}\sup_{\{\theta:|\theta-\theta_0|<\delta\}}\mathbf{E}_{\theta,u}l(\frac{1}{\varepsilon}(\hat{\theta}_\varepsilon-\theta)) \geq \mathbf{E}l(\xi/\sqrt{\lambda_0(\theta_0)}). \tag{12}$$

So we see that u_0 is the optimal input for the parameter estimation and λ_0 is Fisher information for this input.

At the end of this section we consider the following simple example.

Let $L(x)$ be a self-adjoint elliptic operator in D with zero boundary conditions on ∂D, and $L(x,\theta) = \frac{1}{\theta}L(x)$. Let $\Theta = \mathbf{R}^1 \setminus \{0\}$ and K be an identity operator. Then $S(x,\theta) = \theta L^{-1}u(x)$, $S'_\theta(x,\theta) = L^{-1}u(x)$, Fisher information $I(\theta,u)$ does not depend on θ and can be written in the form

$$I(\theta,u) = \|L^{-1}u(x)\|^2$$

So the problem of the input optimization is reduced to the problem of the maximization of $(L^{-2}u,u)$ over $\mathcal{U}$. It is well known, that L^{-1} and L^{-2} are the compact self-adjoint operators in $L_2(D)$. Denote $\lambda_0 > 0$, $u_0(x)$ the minimal eigenvalue and the corresponding eigenfunction of the operator L. Then we can conclude that the optimal input coincides with u_0 and $I(\theta) = \lambda_0{}^{-2}$. The estimator for this example can be written in the form

$$\hat{\theta}_\varepsilon = \lambda_0 \int_D u_0(x)Y_\varepsilon(dx)$$

It is easy to check that $\mathbf{E}(\hat{\theta}_\varepsilon - \theta)^2 = \varepsilon^2\lambda_0{}^2$ and thus $\hat{\theta}_\varepsilon$ is the efficient (not only asymptotically!) estimator of θ for the quadratic loss function.

For this example neither Fisher information nor the optimal input depend on the unknown parameter. In general it is not the case, and the asymptotically efficient estimators and corresponding inputs can be found only adaptively.

In the next section we will formulate some results that were obtained in [15] for the ODE case with a long observation time interval.

3 Optimal Inputs for ODE's

Denote $L(\theta,\lambda) = a_0(\theta)\lambda^n + \ldots + a_{n-1}(\theta) + a_n(\theta)$ a polynomial with coefficients depending smoothly on $\theta \in \Theta$. At this section we consider the case where $L(\theta,x) = \mathcal{L}(\theta,\frac{d}{dx})$, $x \in [0,T]$. Denoting the argument by t instead of x, we obtain the equation (1) in the form

$$\sum_{i=0}^{n} a_i(\theta)S_t^{(n-i)}(t,\theta) = u(t); \quad S^{(k)}(0,\theta) = 0; \quad k = 0,\cdots,n-1. \tag{13}$$

We assume also that the input $u(t) \in \mathbf{R}$ subject to the average energy constraint over a finite time horizon T

$$\frac{1}{T}\int_0^T |u(t)|^2 dt \le 1, \tag{14}$$

Suppose that the observed process $Y_T(t)$ has the stochastic differential

$$dY_T(t) = KS(t,\theta)dt + \sigma dw(t), \ Y_T(0) = 0, t \in [0,T] \tag{15}$$

where S is a solution of (13), K is a linear integral operator

$$KS(t) = \int_0^T K(t-s)S(s)\, ds,$$

$w(t)$ is the standard Wiener process in $\mathbf{R}^1$. We assume that $K(.) \in \mathbf{L}_1(\mathbf{R}^1)$ and σ is a positive constant. The equation (15) models the observation of $KS(t)$ in a WGN of intensity σ^2.

The problems are:

(i) Given the energy constraint (14), determine an optimal input signal $u_T(t)$ that maximizes (or asymptotically maximizes for $T \to \infty$) the Fisher information

$$I_T(u) = \frac{1}{\sigma^2} \int_0^T |K \frac{\partial S}{\partial \theta}(t, \theta))|^2 dt. \tag{16}$$

As a rule, this optimal input signal depends on unknown parameter θ.

(ii) To propose the optimal or asymptotically optimal input signal based on $Y_T(t)$ only and the asymptotically optimal estimator.

Denote

$$\mathcal{L}'_\theta(\theta, \lambda) = \frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \theta} = \sum_{i=0}^{n} a'_i(\theta) \lambda^{n-i}; \ \hat{K}(z) = \int_{\mathbf{R}} e^{isz} K(s)\, ds.$$

It was proven in [15] that the structure of optimal inputs is essentially different in the stable, neutral and unstable cases. For the stable case when, for any $\theta \in \Theta$

$$\Re e \ \mathrm{Spec}\, \mathcal{L}(\theta) < 0, \tag{17}$$

(all roots of the polynomial $\mathcal{L}(\theta, \lambda)$ have a negative real part) the following theorem is valid.

Theorem 2 *Suppose that the equation (13) is stable in the sense (17). Then*

$$\liminf_{T \to \infty} \frac{1}{T} \sup_{u \in \mathcal{U}} I_T(u, \theta) = \sup_{z \in \mathbf{R}} \Big\{ \Big| \frac{\mathcal{L}'_\theta(\theta, iz)}{\mathcal{L}^2(\theta, iz)} \Big|^2 |\hat{K}(z)|^2 \Big\} := I_{\max}(\theta). \tag{18}$$

The corresponding asymptotically optimal input has the form $\exp\{iz_0 t\}$, *where* z_0 *is a value of argument maximizing the right-hand-side in (18).*

As it was mentioned above, the optimal input depends on unknown parameter θ and therefore cannot be implemented directly. But under natural additional assumptions asymptotically optimal optimal input and the corresponding asymptotically efficient estimator can be found for this problem.

Assume that there exist a constant $c > 0$ such that for any $\theta \in \Theta$ and $z \in \mathbf{R}^1$

$$|K(\mathcal{L}(\theta + h, iz) - \mathcal{L}(\theta, iz))| \geq c|h| \tag{19}$$

and the functions $\mathcal{L}(\theta, iz), \mathcal{L}'_\theta(\theta, iz)$ are uniformly continuous in θ for any z.

Consider the following two-steps procedure for the implementation of an asymptotically optimal input. On the time interval $[0, T^{-\alpha}]$, with $\alpha > 0$ small enough, the input signal is a constant $\kappa = T^{1/2}$. Under the condition (19) it follows from the Theorem 3.5.1 in [6] that the maximum likelihood estimator (MLE) $\theta_T^{(1)}$ for θ based on $Y_T(t)$, $t \in [0, T^{-\alpha}]$, is consistent for θ. Let now Z_T be one of real numbers maximizing the expression in the right-hand-side of (18) for $\theta = \theta_T^{(1)}$. Then choosing the input $\exp\{iZ_T t\}$ for $t \in [T^{-\alpha}, T]$ it is possible to prove, using again Theorem 3.5.1 in [6], that MLE $\theta_T^{(2)}$ based on $Y_T(t)$,$t \in [T^{-\alpha}, T]$ is asymptotically efficient in the sense

$$\lim_{T \to \infty} \mathbf{E} |\sqrt{T}(\theta_T^{(2)} - \theta)|^2 = [I_{\max}(\theta)]^{-1}.$$

Instead of using MLE on the second step, sometimes it is possible to use the Fisher's adaptive score correction method , starting from $\theta_T^{(1)}$, see [5, 2, 15].

4 Optimal Input Design for Parabolic Equations

4.1 Estimation of Diffusion Coefficient

Consider the input design problem for the estimation of the diffusion coefficient for the parabolic equation

$$\frac{\partial S}{\partial t} = \frac{1}{\theta}\frac{\partial^2 S}{\partial x^2} + u(t,x);\ S(0,x) = 0. \tag{20}$$

Here the pair (t,x) is in the cylinder $\mathcal{C}_T = [0,T] \times \mathbf{I}$, $\mathbf{I}$ is the interval $[0, 2\pi]$ with the identified end points (circle). We assume that the input signal $u(t,x)$ has the average energy constraint

$$\|u\|^2 := \int_0^T \int_0^{2\pi} u^2(t,x)dx\ dt \leq 2\pi T, \tag{21}$$

and $\theta \in \Theta \subset \mathbf{R} \setminus \{0\}$ (We consider θ^{-1} at the right-hand-side of (20) for the convenience only; it is well known that having the asymptotically efficient estimator (AEE) $\hat{\theta}$ for θ one can often obtain the AEE for $f(\theta)$ of the form $f(\hat{\theta})$).

Similar to (15), we assume that the observed random measure $Y(dt \times dx)$ can be written as:

$$Y(dt \times dx) = S(t,x,u)dtdx + \sigma W(dt \times dx);\ \ (t,x) \in \mathcal{C}_T \tag{22}$$

where $W(dt \times dx)$ is a generalized orthogonal Gaussian random measure so that for any time intervals from $[0,T]$ and Borel sets Γ_1, Γ_2

$$\mathbf{E}W([0,t] \times \Gamma) = 0$$
$$\mathbf{E}W([0,t_1] \times \Gamma_1)W([0,t_2] \times \Gamma_2) = (t_1 \wedge t_2)\Lambda(\Gamma_1 \cap \Gamma_2),$$

where Λ is Lebesgue measure on $[0, 2\pi]$. As before, we interpret (22) in a generalized sense (4).

It is well known (see e.g. [4]) that symbolically this random measure can be written as

$$W(dt \times dx) = \sum_n e^{inx} dW_n(t)dx, \tag{23}$$

where $W_j(t)$ are independent standard Wiener processes. Using the Fourier expansion

$$S(t,x) = \sum_{n=-\infty}^{+\infty} S_n(t,\theta)e^{inx}$$

and the formal expansion

$$Y(dt \times dx) = \sum_{n=-\infty}^{+\infty} e^{inx} dY_n(t)dx, \tag{24}$$

we arrive at the following equations for the Fourier components of the observed measure:

$$dY_n(t) = S_n(t,\theta)dt + \sigma dW_n(t),\ \ t \in [0,T],\ n = 0, \pm 1, \pm 2, ... \tag{25}$$

Taking also the Fourier expansion in x for the input and using (20) we arrive at the following equations for $S_n(t,\theta)$

$$\frac{dS_n}{dt} = -\frac{n^2}{\theta}S_n + u_n(t);\ \ S_n(0,\theta) = 0;\ \ n = 0, \pm 1, \pm 2, ... \tag{26}$$

Note that the component $S_0(t)$ is not informative because it does not depend on θ.

Denote $I_n(\theta, T, u_n)$ the Fisher information for the model (26),(25), and assume that the input constraint has the form

$$\int_0^T |u_n(t)|^2 d\,t \leq 2\pi |A_n|^2 T \tag{27}$$

Then the Fisher information for the problem for (20)- (22) is

$$I(\theta, T, u) = \sum_{n=-\infty}^{+\infty} I_n(\theta, T, u_n)$$

(recall that the observation components $Y_n(t)$ are independent) and the optimal input design problem can be reduced to the problem of a choice $u_n(t)$ and A_n maximizing $\sum I_n(\theta, T, u_n)$ over (27) under the constraint

$$\sum |A_n|^2 \leq 1. \tag{28}$$

It follows from section 3 that for the fixed A_n asymptotically the best input for the problem (25)-(27) has the form $u_n(t) = A_n e^{iz_n t}$ and the corresponding Fisher information has the form

$$I_n(\theta, T) = \frac{2\pi T |A_n|^2 n^4}{\sigma^2 (|z_n|^2 |\theta|^2 + n^4)^2}.$$

It is clear from this expression that

$$\limsup_{T\to\infty} \sup_{z_n, A_n} \frac{1}{T} \sum_n I_n(\theta, T) = 2\pi/\sigma^2$$

and

$$u(t, x) = B_1 \cos x + B_2 \sin x, \quad B_1^2 + B_2^2 = 1 \tag{29}$$

is the corresponding optimal input signal. Note also that $z_n = 0$ for all n and $A_n = 0$ for $n \neq \pm 1$. Consider the estimator

$$\hat{\theta}_T = \frac{1}{2\pi T} \int_0^T \int_0^{2\pi} (B_1 \cos x + B_2 \sin x) Y(dt \times dx) \tag{30}$$

It is easy to check that $\hat{\theta}_T$ is the asymptotically efficient estimator in the sense

$$\mathbf{E}(\hat{\theta}_T - \theta)^2 = \frac{\sigma^2}{2\pi T}(1 + o(1)) \text{ for } \varepsilon \to 0.$$

So we have arrived at the theorem.

Theorem 3 *The asymptotically optimal choice of the input signal for the model (20), (21), (22) can be written in the form (29), and (30) is the corresponding asymptotically efficient estimator in the mean square sense.*

4.2 Estimation of Drift

Consider now the optimal input design problem for the estimation of the drift coefficient for the parabolic equation. The function S for in this case is the solution of a problem

$$\frac{\partial S}{\partial t} = \frac{\partial^2 S}{\partial x^2} + \theta \frac{\partial S}{\partial x} + u(t,x); \ S(0,x) = 0. \tag{31}$$

We assume again that the input signal $u(t,x)$ has the average energy constraint (21) and the equation (31) is on the cylinder C_T. We assume also that the observed random measure $Y(dt \times dx)$ has the form (22) with the generalized orthogonal Gaussian random measure $W(dt \times dx)$ from the previous subsection.

Then, considering the Fourier expansion for S, W, u and Y, we arrive at the equation (25) for the Fourier components of the observed measure, equation (27) for the components of the input (A_n subject to the constraint (28) again) and the equation

$$\frac{dS_n}{dt} = -n^2 S_n + in\theta S_n + u_n(t); \ \ S_n(0,\theta) = 0; \ \ n = 0, \pm 1, \pm 2, ... \tag{32}$$

for the Fourier components of S. Using again the results of section 3 we can write asymptotically the best input for the n-th component of the form $u_n(t) = A_n e^{iz_n t}$ and the corresponding Fisher information is equal to

$$I_n(\theta, T) = \frac{2\pi T |A_n|^2 n^2}{\sigma^2((z_n - \theta n)^2 + n^4)^2}.$$

So we arrive at the maximization problem

$$\max_{z_n, A_n} T \sum_n \frac{|A_n|^2 n^2}{\sigma^2((z_n - \theta n)^2 + n^4)^2} \tag{33}$$

under the constraints (28) and $z_n \in \mathbf{R}^1$. It is easy to see that the maximum in (33) is attained for $z_n = \theta n, A_n = 0$ for $n \neq \pm 1$, and thus

$$\lim_{T \to \infty} \max \frac{I(T, \theta, u)}{T} = \frac{2\pi}{\sigma^2}. \tag{34}$$

The asymptotically optimal input for this problem can be found in the form

$$u(t,x) = A\cos(\theta t + x) + B\sin(\theta t + x), \quad A^2 + B^2 = 2.$$

But for this case the optimal input depends on unknown parameter θ and cannot be implemented directly. So at the beginning let us try to use the input signal of the form $u(t,x) = \sqrt{2}\cos(\omega t + x)$, $\omega \in \mathbf{R}$ is arbitrary. It is easy to see that the solution of the problem (31) for this input signal can be written in the form

$$S(t,x,\theta) = \frac{\sqrt{2}}{(\omega - \theta)^2 + 1}\{\cos(\omega t + x) + (\omega - \theta)\sin(\omega t + x)\} + \tilde{S}(t,x,\theta). \tag{35}$$

Here and below we will denote by $\tilde{S}(t,x,\theta)$ any function which converges to 0 exponentially fast as $t \to \infty$ together with its partial derivative with respect to θ. It is easy to obtain from (35) that the Fisher information for the estimation problem (31),(22) has the behavior

$$\frac{1}{\sigma^2}\int_0^{2\pi}\int_0^T (\frac{\partial S}{\partial \theta})^2 dt\, dx = \frac{2\pi T}{\sigma^2}\Big(\frac{1}{(\omega - \theta)^2 + 1} + o(1)\Big), \quad T \to \infty. \tag{36}$$

It follows from (36) and (34) that the asymptotic efficiency of the best estimator based on this input signal is equal to $[(\omega-\theta)^2+1]^{-1}$, so it is less than 1.

It is possible to construct asymptotically efficient estimator also with help of the following two-step procedure analogous to one described at the section 3. We can choose, for instance,

$$u(t,x)=\cos x, \quad t\in[0,\sqrt{T}]. \tag{37}$$

Then we can use the MLE $\hat{\theta}_T^{(1)}$ of θ on the base of (22) for $t\in[0,\sqrt{T}]$. Using the results of [6] (as it was proven above this estimation problem can be reduced to the estimation problem for one-dimensional WGN, which has been considered in [6]) we can assert, that $\hat{\theta}_T^{(1)}$ is the consistent estimator of θ. On the interval $[\sqrt{T},T]$ we can choose the input

$$u(t,x)=\sqrt{2}\cos(\hat{\theta}_T^{(1)}t+x), \quad t\in[\sqrt{T},T]. \tag{38}$$

Then (see (35)) the solution of (31) can be written at the form

$$S(\theta,t,x)=\frac{\sqrt{2}}{(\hat{\theta}_T^{(1)}-\theta)^2+1}\left\{\cos(\hat{\theta}_T^{(1)}t+x)+(\hat{\theta}_T^{(1)}-\theta)\sin(\hat{\theta}_T^{(1)}t+x)\right\}+\tilde{S}(t,x,\theta). \tag{39}$$

It is not hard to check using the expressions (35),(39),that the Fisher information for the input (37),(38) has the asymptotic behavior $T/\sigma^2(1+o(1))$ for $T\to\infty$. Using again the results of [6] we can assert that the MLE $\hat{\theta}_T^{(2)}$ for the observation of the signal (4) in the WGN of intensity σ (see(22)) on the interval $[\sqrt{T},T]$ is asymptotically efficient for the quadratic loss function. It means that

$$\lim_{T\to\infty} T\mathbf{E}(\hat{\theta}_T^{(2)}-\theta)^2=\sigma^2/2\pi. \tag{40}$$

So we arrive at the theorem.

Theorem 4 *An asymptotically optimal input for the estimation of drift for the model (31), (21),(22) can be found in the form (37), (38). The estimator $\hat{\theta}_T^{(2)}$ constructed with help of this input and the two-step estimation procedure described above is asymptotically efficient for the quadratic loss function.*

Remark. Analogous to the ODE case it is possible to use score correction method at the second step instead of MLE , see [5, 2, 15]. We shall consider this approach in another paper.

References

[1] Aoki, H., Staley, R.M., On input signal synthesis in parameter identification, *Automatica*, **6** (1970), pp. 431–440.

[2] Bickel P., Klaassen Y., Ritov Y., Wellner J., *Efficient and Adaptive Estimation for Semiparametric Models*, John Hopkins Univ. Press, Baltimore, 1993.

[3] Gikhman, I.I. and Skorohod, A. V., *The theory of stochastic processes*, I [Fundamental Principles of Mathematical Sciences], 210. Springer-Verlag, Berlin-New York, 1980. viii+574 pp.

[4] Da Prato G., Zabczyk J., *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, 1992.

[5] Golubev G., Fisher's method of scoring in a frequency estimation problem. *Proceedings of LBMI Seminars.* Leningrad, 1982, **108**, 22–45.

[6] Ibragimov, I.A., Khasminskii, R.Z., *Statistical estimation. Asymptotic theory,* Springer Verlag, New York – Heidelberg – Berlin, 1981.

[7] Hida T., Kuo H.-H., Potthoff J., Streit L., *White Noise*, Kluwer,Dordrecht–Boston–London, 1993.

[8] Kalaba, R.E., Springarn, K., Optimal inputs and sensitivities for parameter estimation, *Jour. Optim. Theory and applications*, **11** (1973), pp. 56–67.

[9] Kalaba, R.E., Springarn, K., *Control, Identification and Input Optimization*, Plenum Press, New-York, 1982.

[10] Levadi, V.S., Design of input signals for parameter estimation, *IEEE Trans. on Automatic Control*, **11** (1966), pp. 205–211.

[11] Levin, M.J., Optimal estimation of impulse response in the presence of noise, *IRE Trans. on Circuit Theory*, **7** (1960), pp. 50–56.

[12] Mehra, R.K., Optimal inputs for linear system identification, *IEEE Trans. on Automatic Control*, **AC–19** (1974), pp. 192–200.

[13] Mehra, R.K., Optimal input signals for parameter identification in dynamical systems – Survey and new results, *IEEE Trans. on Automatic Control*, **AC–19** (1974), pp. 753–768.

[14] Mehra, R.K., Lainiotis, D.G., *System Identification: Advances and Case Studies*, Academic Press, New-York, 1976

[15] Ovseevich, A.I., Khasminskii, R.Z., Chow, P.-L.,Adaptive experimental design for parameter estimation in linear systems,*Problems of Information Transmission*, 36, 2, (2000), pp.38-68.

Linear Cumulant Control and Its Relationship to Risk-Sensitive Control

Libin Mou*	Stanley R. Liberty	Khanh D. Pham	Michael K. Sain[†]
Dept. of Mathematics	Office of the Provost	Dept. Electrical Eng.	Dept. Electrical Eng.
Bradley University	Bradley University	Univ. of Notre Dame	Univ. of Notre Dame
Peoria, IL 61625	Peoria, IL 61625	Notre Dame, IN 46556	Notre Dame, IN 46556

Abstract

Matrix differential equation descriptions of the cumulants of an integral quadratic cost associated with a linear system with white-noise input were derived in the mid-70s using generalized Karhunen-Loeve expansion techniques. Here, these same descriptions are derived directly from the cumulant generating function of the cost. A generalization of the k-cumulant control problem class introduced in 1998 is also presented. The solution to this more general class of optimal cumulant control problems is given, and the risk sensitive control problem of optimizing the cumulant generating function of the LQG cost is shown to be included in this cumulant control class.

1 Introduction

In 1998 Pham, Liberty and Sain introduced a general class of Linear-Quadratic-Gaussian (LQG) control problems in which the objective is minimization of a performance index that is a finite, linear combination of cumulants of integral quadratic cost over linear, memoryless, full-state-feedback control laws [9]. The formulation of this "k-cumulant" optimization problem utilizes a coupled matrix differential equation description of the cumulants of an integral quadratic form (IQF) in a Gaussian process. These equations were first derived by Liberty and Hartwig [8] in 1976. In that work the Gaussian process is the state of a linear system with white-noise input. The work reported in [8] evolved from that of Liberty [7] in 1971 where it was observed that all cumulants of such IQFs are quadratic-affine in the mean of the system initial state. Two examples of control problems in the k-cumulant class for $k = 1$ and $k = 2$ respectively are the classical minimum mean LQG problem and the Minimum Cost Variance (MCV) problem [12, 13].

Since the early 1970s other researchers have developed the theory of risk-sensitive control; see for example [2, 3, 4, 5, 6, 14]. Also see [16] where Won, Sain and Spencer presented a brief history of risk-sensitive control and pointed out the close relationship between risk-sensitive control and "the notion of optimal cost cumulants, in particular cost variance." In this paper we demonstrate an even deeper relationship between risk-sensitive and cumulant control.

---

*The first author wishes to acknowledge the support of a Caterpillar Fellowship. [†]The work of this author is supported by the Frank M. Freimann Chair in Electrical Engineering, University of Notre Dame, and by the National Science Foundation under Grants CMS 93-01584, CMS 95-00301, CMS 95-28083 and CMS 99-00234.

In the next section we demonstrate that the coupled matrix differential equations presented in [8] and utilized in [9] can be generated directly from the well known quadratic-affine form of the cumulant generating function of the (IQF) cost. In section 3 we consider a further generalization of k-cumulant control, which we simply refer to as cumulant control. Observed formally by Won [15], the generalization consists of replacing the finite linear combination of cost cumulants in the performance index with an infinite series of cost cumulants. The solution to this "cumulant control" problem is introduced in section 3. Finally, in section 4 we demonstrate the relationship between the solution to the risk-sensitive control problem (for the linear, memoryless, full-state-feedback case) and the solution to the cumulant control problem.

2 Cumulant Descriptions

Consider the linear system

$$dx = Fxds + Gdw,\ x\left(t_0\right) = x_0 \tag{1}$$

where the matrix functions $F\left(s\right) \in R^{n \times n}$, $G\left(s\right) \in R^{n \times p}$ are continuous, $w\left(s\right) \in R^p$ is the standard Wiener process, and x_0 is non-random. For every pair $\left(t_0, x_0\right)$, associate with (1) a cost

$$J\left(t_0, x_0\right) = x^T\left(t_f\right) Q_f x\left(t_f\right) + \int_{t_0}^{t_f} x^T\left(s\right) N\left(s\right) x\left(s\right) ds \tag{2}$$

where $Q_f \in R^{n \times n}$ is a constant matrix and ≥ 0, $N\left(t\right) \in R^{n \times n}$ is a continuous matrix and ≥ 0. Here J is considered to be a function of t_0 and x_0. In the following we will replace $\left(t_0,\ x_0\right)$ by $\left(t, x\right) \in \left[t_0, t_f\right] \times R^n$, and $J\left(t_0, x_0\right)$ by $J\left(t, x\right)$, the "cost-to-go", which is (2) with t_0 replaced by t and $x\left(s\right)$ replaced by the solution of (1) with initial condition $x\left(t\right) = x$.

It is well-known (see [1] for example) that the moment-generating and cumulant-generating functions of the cost-to-go

$$\phi\left(\theta, t, x\right) = E\{e^{\theta J(t,x)}\} \text{ and } \psi\left(\theta, t, x\right) = \ln \phi\left(\theta, t, x\right),$$

where $E\{\cdot\}$ is the expected value of the enclosed random variable, can be expressed in terms of the solution of a particular Riccati equation as follows:

Theorem 1. *For fixed θ, let the functions $\rho\left(\theta, t\right)$ and $S\left(\theta, t\right)$ be solutions to the equations*

$$\begin{cases} S' + F^T S + SF + 2SWS + \theta N = 0 \\ S\left(\theta,\ t_f\right) = \theta Q_f, \end{cases} \tag{3}$$

and

$$\begin{cases} \rho' = -\rho\, tr(SW), \\ \rho\left(\theta,\ t_f\right) = 1, \end{cases} \tag{4}$$

where $W = GG^T$ and (') denotes the derivative. Then

$$\begin{aligned} &(a) \quad \phi\left(\theta, t, x\right) = \rho\left(\theta, t\right) exp\left(x^T S\left(\theta, t\right) x\right) \\ &(b) \quad \psi\left(\theta, t, x\right) = d\left(\theta, t\right) + x^T S\left(\theta, t\right) x \end{aligned}$$

with $d(\theta,t) = \ln \rho(\theta,t)$ *satisfying*

$$\begin{cases} d' = -\,tr\,(SW), \\ d(\theta,\, t_f) = 0. \end{cases}$$

☐

Recall that for $i = 1, 2, \cdots$ the moments $m_i(t,x)$ of $J(t,x)$ are defined as the coefficients of $\theta^i/i!$ in the McLaurin series expansion of the moment-generating function $E\{e^{\theta J}\}$, while the cumulants $\kappa_i(t,x)$ of $J(t,x)$ are the coefficients of $\theta^i/i!$ in the McLaurin series expansion of the cumulant-generating function

$$\ln E\{e^{\theta J}\} = \sum_{i=1}^{\infty} \frac{\theta^i}{i!}\kappa_i. \tag{5}$$

We now present an alternative and more direct derivation of the cumulant expressions contained in [8].

Theorem 2. *The cumulants* $\kappa_i(t,x)$ *of* $J(t,x)$ *can be expressed as*

$$\kappa_i(t,x) = d_i(t) + x^T H_i(t)\, x$$

where d_i *and* H_i *satisfy the following Lyapunov-type differential equations:*

$$H_1' + F^T H_1 + H_1 F + N = 0,\ H_1(t_f) = Q_f,$$

$$H_i' + F^T H_i + H_i F + 2\sum_{j=1}^{i-1} \frac{i!}{j!\,(i-j)!} H_j W H_{i-j} = 0,\ H_i(t_f) = 0,\ i \geq 2,$$

$$d_i' = -tr\,(H_i(t)\, W),\ d_i(t_f) = 0,\ i \geq 1.$$

☐

Proof. Utilizing the functions in Theorem 1 define

$$H_i(t) = \frac{\partial^i}{\partial \theta^i} S(0,t) \text{ and } d_i(t) = \frac{\partial^i}{\partial \theta^i}\rho(0,t).$$

Then by expanding the cumulant generating function expression of Theorem 1.b in a McLaurin series we get

$$\psi(\theta,t,x) = \rho(\theta,t) + x^T(t)\, S(\theta,t)\, x = \sum_{i=1}^{\infty} \left(d_i(t) + x^T H_i(t)\, x\right)\frac{\theta^i}{i!}.$$

Matching coefficients with the series in (5) we observe that the i^{th} cumulant of $J(t,x)$ can be expressed as

$$\kappa_i(t,x) = d_i(t) + x^T H_i(t)\, x.$$

The equation for H_1 follows by evaluating the $\frac{\partial}{\partial \theta}$ of (3) at $\theta = 0$ and using the fact $S(0,t) = 0$. For $i \geq 2$, by evaluating the $\frac{\partial^i}{\partial \theta^i}$ of (3) at $\theta = 0$ and using $S(0,t) = 0$, we derive the differential equations and boundary conditions satisfied by $H_i(t) = S^{(i)}(0,t)$. Similarly from (4) we get the equations $d_i' = -\mathrm{tr}(H_i(t)\, W)$ and $d_i(t_f) = 0$. ☐

3 Cumulant Control Problems

We now consider a general cumulant control problem associated with the following linear control system and integral quadratic cost function:

$$dx_c = (Ax_c + Bu)\,ds + G dw,\ s \in [t_0, t_f]\,;\ x_c(t_0) = x_0, \tag{6}$$

$$J_c = x_c^T(t_f)\,Q_f x_c(t_f) + \int_{t_0}^{t_f} \left(x_c^T Q x_c + u^T R u\right) ds \tag{7}$$

where the matrix functions $A(t) \in R^{n\times n}$, $B(t) \in R^{n\times m}$, $G(t) \in R^{n\times p}$ are continuous. $Q_f \in R^{n\times n}$ is constant matrix and ≥ 0, $Q(t) \in R^{n\times n}$ and $R(t) \in R^{m\times m}$ are continuous on $[t_0, t_f]$, $Q(t) \geq 0$ and $R(t) > 0$. We will assume that u is a linear memoryless, full-state-feedback control given by

$$u(s) = K(s)\,x(s),$$

with feedback gain $K(s) \in R^{m\times n}$. Then the system (6) and the cost (7) can be written as

$$dx_c = Fx_c ds + G dw,\ x_c(t_0) = x_0, \tag{8}$$

$$J_c(t_0, x_0, K) = x_c^T(t_f)\,Q_f x_c(t_f) + \int_{t_0}^{t_f} x_c^T(s)\,N(s)\,x_c(s)\,ds, \tag{9}$$

where

$$F(s) = A(s) + B(s)\,K(s)\,,\ N(s) = Q(s) + K^T(s)\,R(s)\,K(s)\,. \tag{10}$$

It is well-known that cumulants and moments are expressible in terms of each other, however, cumulants are more useful quantities in stochastic optimal control because of their common quadratic-affine form.

For a sequence of real numbers $\mu = \{\mu_1, \mu_2, \cdots\}$, we can use (10) and the cumulant expressions in Theorem 2 to define a series of cumulants

$$\kappa(t_0, x_0) = x_0^T \sum_{i=1}^{\infty} \mu_i H_i(t_0) x_0 + \sum_{i=1}^{\infty} \mu_i d_i(t_0) \tag{11}$$

Without loss of generality, we will assume that $\mu_1 = 1$.

Define $\mathcal{H}(s) = (H_1(s)\,, H_2(s)\,, \cdots)$ and $\mathcal{D}(s) = (d_1(s)\,, d_2(s)\,, \cdots)$ and rewrite the equations in Theorem 2, for H_i and d_i as

$$\begin{cases} H_1' + \mathcal{F}_1(\mathcal{H}, K) = 0,\ H_1(t_f) = Q_f, \\ H_i' + \mathcal{F}_i(\mathcal{H}, K) = 0,\ H_i(t_f) = 0,\ i \geq 2, \\ d_i' + \mathcal{G}_i(\mathcal{H}) = 0,\ d_i(t_f) = 0,\ i \geq 1. \end{cases} \tag{12}$$

where

$$\begin{cases} \mathcal{F}_1(\mathcal{H}, K) = F^T H_1 + H_1 F + N, \\ \mathcal{F}_i(\mathcal{H}, K) = F^T H_i + H_i F + 2\sum_{j=1}^{i-1} \frac{i!}{j!(i-j)!} H_j W H_{i-j}, i \geq 2, \\ \mathcal{G}_i(\mathcal{H}) = \operatorname{tr}(H_i(s)\,W)\,. \end{cases} \tag{13}$$

Now let

$$\Phi(t_0, \mathcal{H}(t_0), \mathcal{D}(t_0); x_0; \mu) = \kappa(t_0, x_0).$$

The cumulant control problem that we wish to solve is

$$\underset{K \in \mathcal{K}_\mu}{Min} \quad \Phi(t_0, \mathcal{H}(t_0), \mathcal{D}(t_0); x_0; \mu)$$

subject to (12) and (13). The class, $\mathcal{K}_\mu$ of admissible feedback gains is the set of all $m \times n$ matrix functions that are continuous on $[t_0, t_f]$ with values in a compact subset of the vector space of all $m \times n$ real matrices, and which yield existence of solutions on $[t_0, t_f]$ to (12) and in combination with the given μ_i's yield a convergent series in (11).

Remark: *For a given stochastic linear system are there μ-sequences that result in a rich class K_μ of admissible function gains? The answer to this question is affirmative. Due to space limitation here we present some results without proof. It can be shown that there is a $\sigma > 0$ such that if $\{\mu_i^*\}_{i=1}^{\infty}$ where $\mu_i^* = i!\mu_i$ satisfies*

$$0 \leq \mu_i^* \leq \sigma \mu_j^* \mu_{i-j}^* \tag{14}$$

for $i = 2, \ldots$ and $j = 1, \ldots, i-1$, then K_μ contains many admissible feedback gains (for example, it contains all K's with small norm) with convergence of the series (11).

To illustrate further that there are many μ_i^*'s satisfying (14), consider the following. For a given $\sigma > 0$, there are intervals $\{[a_i, b_i]\}_{i=1}^{\infty}$ such that if $\{\mu_i^*\}_{i=1}^{\infty}$ is a sequence with $\mu_i^* \in [a_i, b_i]$, then it satisfies the condition (14). For example, let $a_i = \alpha\theta^{i-1}$ and $b_i = \sigma\alpha^2\theta^{i-2}$ for some numbers $\alpha, \theta > 0$ such that $\sigma\alpha > \theta$ and $\mu_i^* \in [a_i, b_i]$. Then for $i = 2, 3, \cdots$ and $j = 1, \cdots, i-1$,

$$\frac{\mu_i^*}{\mu_j^* \mu_{i-j}^*} \leq \frac{b_i}{a_j a_{i-j}} = \sigma.$$

A particular example satisfying (14) with $\sigma = 1$ is $\mu_i = \theta^i / i!$, which is the risk-sensitive case; see (5).

We prove

Theorem 3. *The admissible feedback gain K that minimizes $\Phi(t_0, \mathcal{H}(t_0), \mathcal{D}(t_0); x_0; \mu)$ and the corresponding $H_1, H_2, \ldots$ and $d_1, d_2, \ldots$ satisfy the following equations*

$$\begin{cases} F = A + BK, \\ K = -R^{-1}B^T \sum_{i=1}^{\infty} \mu_i H_i, \\ H_1' + F^T H_1 + H_1 F + Q + K^T R K = 0, \ H_1(t_f) = Q_f, \\ H_i' + F^T H_i + H_i F + 2\sum_{j=1}^{i-1} \frac{i!}{j!(i-j)!} H_j W H_{i-j} = 0, \ H_i(t_f) = 0, \ i \geq 2, \\ d_i' = -tr(H_i(s) W), \ d_i(t_f) = 0, \ i \geq 1. \end{cases}$$

$\square$

Proof. The control problem is a standard Mayer type problem.

For $t \in [t_0, t_f]$, consider (13) on the interval $[t_0, t]$ with terminal values $H_i(t) = y_i$ and $d_i(t) = z_i$. From the representation of κ, we look for a value function of the following form

$$V(t, y) = x_0^T \sum_{i=1}^{\infty} \mu_i (y_i + h_i(t)) x_0 + \sum_{i=1}^{\infty} \mu_i (z_i + e_i(t))$$

where $h_i(t)$'s and $e_i(t)$'s are functions to be determined. By the Verification Theorem [4, Thms 4.1, 4.4], $V(t,y)$ has to satisfy

$$\begin{matrix} Min \\ K \end{matrix} \left\{ -\frac{\partial V}{\partial t} + \sum_{i=1}^{\infty} \frac{\partial V}{\partial y_i} \mathcal{F}_i + \sum_{i=1}^{\infty} \frac{\partial V}{\partial z_i} \mathcal{G}_i \right\} = 0. \tag{15}$$

Substituting (13) into (15) we obtain

$$\begin{matrix} Min \\ K \end{matrix} \left\{ x_0^T \left(-\sum_{i=1}^{\infty} \mu_i h_i'(t) + \sum_{i=1}^{\infty} \mu_i \mathcal{F}_i(y,K) \right) x_0 + \sum_{i=1}^{\infty} \mu_i \left(-e_i'(t) + \mathcal{G}_i(y)\right) \right\} = 0 \tag{16}$$

Note that among all of the terms in $\{\cdots\}$, only $\sum_{i=1}^{\infty} \mu_i \mathcal{F}_i(y,K)$ depends on K. We have

$$\begin{aligned} \sum_{i=1}^{\infty} \mu_i \mathcal{F}_i(y,K) = {} & (A+BK)^T \left(\sum_{i=1}^{\infty} \mu_i h_i \right) + \left(\sum_{i=1}^{\infty} \mu_i h_i \right) (A+BK) + Q + \\ & K^T R K + 2 \sum_{i=2}^{\infty} \mu_i \sum_{j=1}^{i-1} \frac{i!}{j!\,(i-j)!} H_j W H_{i-j} \end{aligned}$$

This matrix quadratic is minimal when $K = -R^{-1} B^T \sum_{i=1}^{\infty} \mu_i h_i$. With this choice of K, let $H_i(t)$ and $d_i(t)$ be the solutions to (13) with $H_i(t) = y_i$ and $d_i(t) = z_i$. Take

$$h_i(t) = H_i(t_0) - H_i(t), \; e_i(t) = d_i(t_0) - d_i(t),$$

then the minimum in (16) will be zero. It turns out that the value function $V(t,y)$ is

$$V(t,y) = x_0^T \sum_{i=1}^{\infty} \mu_i H_i(t_0) x_0 + \sum_{i=1}^{\infty} \mu_i d_i(t_0),$$

which is the optimal value of $\Phi(t_0, \mathcal{H}(t_0), \mathcal{D}(t_0); x_0; \mu)$. □

4 Relationship Between Linear Cumulant Control and Risk-Sensitive Control

The existence of optimal risk-sensitive control is well-known [5, 14]. Denote the optimal value

$$\Psi(\theta, t, x) = \begin{matrix} Min \\ K \in \mathcal{K}_\mu \end{matrix} \left[\ln E\{e^{\theta J_c(t,x,K)}\} \right],$$

where $\theta > 0$ is assumed.

Theorem 4. *The value function $\Psi(\theta, t, x)$ can be expressed as*

$$\Psi(\theta, t, x) = x^T S(\theta, t) x + d(\theta, t),$$

where $S(\theta, t)$ *and* $d(t, \theta)$ *satisfy the Riccati system*

$$\begin{cases} K = -(\theta R)^{-1} B^T S, \\ S' + (A + BK)^T S + S(A + BK) + 2SWS + \theta\left(Q + K^T RK\right) = 0, \\ S(\theta, t_f) = \theta Q_f, \\ d' = -tr(SW),\ d(\theta, t_f) = 0. \end{cases} \tag{17}$$

□

We answer a long-standing question about the relationship between risk-sensitive and cumulant control.

Theorem 5. *Suppose* A, B, Q, R, W *are given matrices as in Theorem 4. Suppose the* K, $H_1 \ldots, H_n, \ldots,$ *and the constant* θ *satisfy the following system*

$$\begin{cases} H_1' + (A + BK)^T H_1 + H_1(A + BK) + Q + K^T RK = 0,\ H_1(t_f) = Q_f, \\ H_i' + (A + BK)^T H_i + H_i(A + BK) + 2\sum_{j=1}^{i-1} \frac{i!}{j!(i-j)!} H_j W H_{i-j} = 0, \\ H_i(t_f) = 0,\ i \geq 2, \\ K = -(\theta R)^{-1} B^T \sum_{i=1}^{\infty} \frac{\theta^i}{i!} H_i. \end{cases} \tag{18}$$

Then the matrix

$$\mathcal{H} = \sum_{i=1}^{\infty} \frac{\theta^i}{i!} H_i$$

satisfies the matrix equation (17)

$$\mathcal{H}' + (A + BK)^T \mathcal{H} + \mathcal{H}(A + BK) + \theta\left(Q + K^T RK\right) + 2\mathcal{H}W\mathcal{H} = 0 \tag{19}$$

and K *is a solution to*

$$\Psi(\theta, t, x) = \underset{K}{Min} \left[ln\ E_{tx} e^{\theta J(t,x,K)}\right],$$

where

$$J(t, x, K) = \int_t^{t_f} \left(x^T Qx + u^T Ru\right) ds + x^T(t_f) Q_f x(t_f).$$

In other words, $\mathcal{H}$ *is the solution to the risk-sensitive control problem.* □

Proof. Multiply the i-th equation in (18) by $\frac{\theta^i}{i!}$ for $i \geq 1$ and add them up. We get

$$\mathcal{H}' + (A + BK)^T \sum_{i=1}^{\infty} \frac{\theta^i}{i!} H_i + \sum_{i=1}^{\infty} \frac{\theta^i}{i!} H_i (A + BK)$$
$$+ \theta\left(Q + K^T RK\right) + 2\sum_{i=2}^{\infty}\sum_{j=1}^{i-1} \frac{\theta^j}{j!} H_j W \frac{\theta^{i-j}}{(i-j)!} H_{i-j} = 0.$$

Since

$$\sum_{i=2}^{\infty}\sum_{j=1}^{i-1} \frac{\theta^j}{j!} H_j W \frac{\theta^{i-j}}{(i-j)!} H_{i-j} = \sum_{j=1}^{\infty} \frac{\theta^j}{j!} H_j W \sum_{k=1}^{\infty} \frac{\theta^k}{k!} H_k = \mathcal{H}W\mathcal{H},$$

we obtain

$$\mathcal{H}' + (A + BK)^T \mathcal{H} + \mathcal{H}(A + BK) + \theta\left(Q + K^T RK\right) + 2\mathcal{H}W\mathcal{H} = 0.$$

Note that this is precisely (19) with $K = -(\theta R)^{-1} B^T \mathcal{H}$. Therefore $\mathcal{H}$ is a solution to the risk-sensitive control problem. □

References

[1] A. Bensoussan, *Stochastic Control of Partially Observable Systems,* London, Cambridge University Press, 1992.

[2] W. H. Fleming and D. Hernandez-Hernandez, "Risk-Sensitive Control of Finite State Machines on an Infinite Horizon II," *SIAM Journal of Control Optimization*, Vol. 37, No. 4, pp. 1048-1069, 1999.

[3] W. H. Fleming and W. M. McEneaney, "Risk-Sensitive Optimal Control and Differential Games", *Lecture Notes in Control and Information Sciences*, Vol. 184, pp. 185–197, Springer-Verlag, Berlin, 1992.

[4] W. H. Fleming and R. W. Rischel, *Deterministic and Stochastic Optimal Control,* Springer-Verlag, 1975.

[5] D. H. Jacobson, "Optimal Stochastic Linear Systems with Exponential Criteria and Their Relation to Deterministic Differential Games," *IEEE Transactions on Automatic Control*, Vol. 18, pp. 124-131, 1973.

[6] M. R. James, "Asymptotic Analysis of Nonlinear Stochastic Risk-Sensitive Control and Differential Games," *Mathematics Control Signals Systems 5,* pp. 401-417, 1992.

[7] S. R. Liberty, "Characteristics Functions of LQG Control," *Ph.D. Dissertation*, Dept. of Electrical Engineering, University of Notre Dame, 1971.

[8] S. R. Liberty and R. C. Hartwig, "On the Essential Quadratic Nature of LQG Control Performance Measure Cumulants", *Information and Control,* Vol. 32, No. 3, pp. 276-305, 1976.

[9] K. D. Pham, S. R. Liberty, and M. K. Sain, "Linear Optimal Cost Cumulant Control: A k-Cumulant Problem Class," *Proceedings Thirty-Sixth Allerton Conference on Communication, Control, and Computing*, pp. 460-469, October 1998.

[10] K. D. Pham, S. R. Liberty, and M. K. Sain, "Evaluating Cumulant Controllers on a Benchmark Structure Protection Problem in the Presence of Classic Earthquakes," *Proceedings Thirty-Seventh Allerton Conference on Communication, Control, and Computing*, pp. 617-626, October 1999.

[11] M. K. Sain and S. R. Liberty, "Performance Measure Densities for a Class of LQG Control Systems," *IEEE Transactions on Automatic Control*, AC-16, No. 5, pp. 431-439, October 1971.

[12] M. K. Sain, "Control of Linear Systems According to the Minimal Variance Criterion: A New Approach to the Disturbance Problem," *IEEE Transactions on Automatic Control*, AC-11, No. 1, pp. 118-122, January 1966.

[13] M. K. Sain, C. H. Won, B. F. Spencer, Jr., and S. R. Liberty, "Cumulants and Risk-Sensitive Control: A Cost Mean and Variance Theory with Application to Seismic Protection of Structures," pp. 427-459 in *Advances in Dynamic Games and Applications*, Annals of the International Society of Dynamic Games, Vol. 5, Jerzy A Filar, Vladimir Gaitsgory, and Koichi Mizukami, Editors. Boston: Birkhauser, 2000.

[14] P. Whittle, *Risk-Sensitive Optimal Control*, New York: John Wiley & Sons, 1990.

[15] C. H. Won, "Cost Cumulants in Risk-Sensitive and Minimal Cost Variance Control," *Ph.D. Dissertation*, Dept. of Electrical Engineering, University of Notre Dame, 1994.

[16] C. H. Won, M. K. Sain, and B. F. Spencer, Jr.,"Relationships in Modern Stochastic Control: Risk-Sensitive and Minimal Cost Variance Control," *Proceedings Allerton Conference on Communication, Control, and Computing*, pp. 343-352, 1996.

REDUCED-ORDER OBSERVER DESIGN FOR UNCERTAIN DISCRETE-TIME STOCHASTIC SYSTEMS WITH GENERAL CRITERIA

Edwin Engin Yaz
Department of Electrical Engineering
University of Arkansas, Fayetteville, AR 72701
E-Mail: eyaz@uark.edu

Yvonne Ilke Yaz
Department of Mathematics
Centenary College, Shreveport, LA 71134
E-Mail: yyaz@centenary.edu

Abstract: A general class of discrete-time uncertain nonlinear stochastic systems with quadratic sum constraints is considered. A linear reduced- order observer for state estimation is presented for various estimation error performance criteria in a unified framework. The observer is of fixed order which is equal to the difference between the state and output vector dimensions. These performance criteria include guaranteed-cost suboptimal versions of estimation objectives like H_2, H_∞, stochastic passivity, etc. The design of fixed-order linear state estimators that satisfy these criteria are given using a common matrix inequality formulation.

Introduction

The problem of fixed(reduced)-order linear observer (state estimator) design is formulated using linear matrix inequalities (LMI) for a general class of uncertain discrete-time nonlinear stochastic systems. Various criteria are used in designing discrete-time estimators leading to a common LMI framework. The reason for choosing this approach is the possible utilization of efficient numerical schemes for solving LMI [1]. In the model used, the system and measurement vectors are assumed to be driven by possibly state-dependent deterministic or stochastic disturbances defined by quadratic sum constraints. These disturbances may be generated e.g. by white noises for which covariance upper bounds are determined by unknown nonlinear functions of the state vector. Such nonlinear models with pointwise quadratic constraints are introduced in [2], and their system theoretic properties are investigated using an LMI approach in [3]. Mean-square stabilizability by state feedback and mean-square detectability concepts for this class of systems are developed and based on these, the infinite horizon linear quadratic regulator is presented in [4]. Covariance assignment formulation is used to characterize static output feedback stabilizability in the mean-square sense and parameterize all stabilizing gains in [5]. Full-order linear minimum variance state estimator design is introduced in [6]. In the present work, various estimation problems including guaranteed-cost suboptimal versions of H_2, H_∞, stochastic passivity, etc. are tackled within a common framework. In that sense, the present work can be viewed as an extension of minimum-variance results of [6] to the case of reduced-order estimation for generalized performance criteria with quadratic sum rather than pointwise quadratic constraints in time.

The following notation is used. $\mathbf{Z}^+$ is the set of nonnegative integers. For an $n-$dimensional vector of real elements $x \in \mathbf{R}^n$, $\| x \|$ denotes the 2−norm. $(x^T x)^{1/2}$. For an $n \times n$ symmetric matrix A, $A > 0$ ($A < 0$) and $A \geq 0$ ($A \leq 0$) denote positive (negative) definite and positive (negative) semidefinite A. The Schur complement formula which states the equivalence of $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \geq 0$ and $C > 0$, $A \geq BC^{-1}B^T$ and $A > 0$, $C \geq B^T A^{-1} B$ is used in this work. $\lambda_{\max}(A)$, $\lambda_{\min}(A)$ (for symmetric A) respectively mean the maximum and minimum eigenvalue of A. $A^\dagger$ denotes the unique Moore-Penrose pseudo-inverse of the A matrix. The form of Rayleigh's inequality $\lambda_{min}(A)I \leq A \leq \lambda_{max}(A)I$ for symmetric A is used. $E\{x\}$ and $E_y\{x\}$ respectively denote the expectation of x and its expectation conditional on y. The interlacing property of expectations $E\{E_y\{x\}\} = E\{x\}$ is also used.

The System and Measurement Models

We assume that the signal is generated by the following system and the measurement equations

$$x_{k+1} = Ax_k + Bp_k \tag{1}$$

$$y_k = C_1 x_k + Dp_k \tag{2}$$

where the state $x_k \in \mathbf{R}^n$ and the measurement $y_k \in \mathbf{R}^p$. The initial state x_0 is assumed to have the known mean $E\{x_0\} = \bar{x}_0$, covariance $E\{x_0 {x_0}^T\} = X_0$, and to be uncorrelated with other noise sources. The unknown disturbance p_k, which may be a nonlinear stochastic function of x_k in general, is defined by a bound on its energy as follows:

$$\sum_{k=0}^{N} E\{\| p_k \|^2\} \leq \sum_{k=0}^{N} E\{\| q_k \|^2\} \tag{3}$$

for any $N \in \mathbf{Z}^+$, where

$$q_k = C_q x_k + D_q p_k \tag{4}$$

This description is quite general allowing the designer to treat, for example:

(a) Linear systems with state-multiplicative noises $C_q x_k v_k$ where v_k is a zero mean uncorrelated noise sequence with an uncertain second moment having a known bound.

(b) Nonlinear systems with random sequences whose powers depend on sector-bound nonlinear function of the state $\psi(x_k)v_k$ where $\| \psi(x_k) \| \leq \alpha \| x_k \|$ where the form of the nonlinearity may not be known but the bound α is known.

(c) Nonlinear systems with a random sequence whose power depends on a bounded nonlinear function of the state $\phi(x_k)C_q x_k v_k$ where $\phi(x)$ satisfies $\| \phi(x) \| \leq \alpha$ for all $x \in \mathbf{R}^n$.

(d) Linear systems with state multiplicative disturbances $C_q x_k w_k$ where $w_k \in l_2$ (i.e. $\sum_{k=0}^{\infty} {w_k}^2 < \infty$) or $w_k \in l_\infty$ (i.e. $\sup_k(|w_k|) < \infty$).

(e) Linear systems having other linear systems in the feedback loop, where the feedback system has an H_∞-norm less than or equal to 1 [1, pg. 122].

Other models that fit this description can be found in [2], [3], and [7]. One can see that some of the most important deterministic and stochastic nonlinear models in use today can be treated in this general framework.

Estimator Design

We define the linear state transformation

$$z_k = C_2 x_k \tag{5}$$

where $z_k \in \mathbf{R}^{n-p}$ and C_2 is a full rank matrix such that $\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$ is nonsingular with

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix}^{-1} = (\,\Omega_1 \quad \Omega_2\,) \tag{6}$$

It is always possible to find a C_2 for a given full rank C_1 such that (6) is true. In fact, based on the full rank properties of C_1, and C_2, we can find explicit expressions for Ω_1 and Ω_2 using the results in [8] as follows:

$$\begin{aligned}(\,\Omega_1 \quad \Omega_2\,) &= (C_1{}^{\dagger}C_1 + C_2{}^{\dagger}C_2)^{-1}(\,C_1{}^{\dagger} \quad C_2{}^{\dagger}\,) \\ &= (C_1{}^T(C_1C_1{}^T)^{-1}C_1 + C_2{}^T(C_2C_2{}^T)^{-1}C_2)^{-1}(\,C_1{}^T(C_1C_1{}^T)^{-1} \quad C_2{}^T(C_2C_2{}^T)^{-1}\,)\end{aligned} \tag{7}$$

z_k is one of the intermediate variables in this estimator design process. Note that it follows from (2), (5), and (6) that if we knew what z_k and p_k were, we would be able to find x_k as follows:

$$x_k = \Omega_1(y_k - Dp_k) + \Omega_2 z_k \tag{8}$$

However, p_k is an unknown disturbance and the estimate of z_k must be used in forming the estimate of x_k

$$\hat{x}_k = \Omega_1 y_k + \Omega_2 \hat{z}_k \tag{9}$$

So, to be able to estimate x_k, an estimate of z_k (which is $(n-p)$-dimensional) is sufficient since the measurement y_k is available. This implies that we need to look at the dynamics of y_k and z_k:

$$\begin{aligned}\begin{pmatrix} y_{k+1} \\ z_{k+1} \end{pmatrix} &= \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}(Ax_k + Bp_k) + \begin{pmatrix} D \\ 0 \end{pmatrix} p_{k+1} \\ &= \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}\left[A\,(\,\Omega_1 \quad \Omega_2\,)\begin{pmatrix} y_k - Dp_k \\ z_k \end{pmatrix} + Bp_k\right] + \begin{pmatrix} D \\ 0 \end{pmatrix} p_{k+1}\end{aligned} \tag{10}$$

where we substituted from (1) and (8). We define

$$\nu_k \doteq z_k - K y_k \tag{11}$$

which is the variable that is directly used in forming the estimator. K is called the observer gain. Note that ν_k has the same dimension as z_k and estimating ν_k is equivalent to estimating z_k since y_k is the known measurement vector, or

$$\hat{\nu}_k = \hat{z}_k - K y_k \tag{12}$$

The dynamics of ν_k can be found from the dynamics of z_k and y_k as follows:

$$\begin{aligned}\nu_{k+1} &= (C_2 - KC_1)A \begin{pmatrix} \Omega_1 & \Omega_2 \end{pmatrix} \begin{pmatrix} y_k - Dp_k \\ z_k \end{pmatrix} \\ &\quad + (C_2 - KC_1)Bp_k - KDp_{k+1}\end{aligned} \tag{13}$$

The estimate of ν_k can be constructed as

$$\begin{aligned}\hat{\nu}_{k+1} &= (C_2 - KC_1)A \begin{pmatrix} \Omega_1 & \Omega_2 \end{pmatrix} \begin{pmatrix} y_k \\ \hat{z}_k \end{pmatrix} \\ &= (C_2 - KC_1)A \begin{pmatrix} \Omega_1 & \Omega_2 \end{pmatrix} \begin{pmatrix} y_k \\ \hat{\nu}_k + K y_k \end{pmatrix}\end{aligned} \tag{14}$$

by using (12) and (13). The resulting estimation error $e_k \doteq \nu_k - \hat{\nu}_k$ has the dynamics

$$\begin{aligned}e_{k+1} &= (C_2 - KC_1)A\Omega_2 e_k + (C_2 - KC_1)(B - A\Omega_1 D)p_k - KDp_{k+1} \\ &\doteq A_0 e_k + B_0 p_k - KDp_{k+1}\end{aligned} \tag{15}$$

The following result summarizes the main contribution of this paper:

Theorem 1. Consider the stochastic nonlinear model (1) - (4), the performance output

$$r_k = C_r e_k + D_r p_k \tag{16}$$

and the reduced-order linear state estimator given by (9), (12) and (14). If the following LMI holds

$$\mathcal{G} = \begin{pmatrix} g_{11} & g_{12} & g_{13} & g_{14} & g_{15} \\ g^T{}_{12} & g_{22} & g_{23} & g_{24} & g_{25} \\ g^T{}_{13} & g^T{}_{23} & g_{33} & g_{34} & g_{35} \\ g^T{}_{14} & g^T{}_{24} & g^T{}_{34} & g_{44} & g_{45} \\ g^T{}_{15} & g^T{}_{25} & g^T{}_{35} & g^T{}_{45} & g_{55} \end{pmatrix} \geq 0 \tag{17}$$

where

$$g_{11} = P_1 - A^T P_1 A - \tau_1 C_q{}^T C_q - \tau_2 A^T C_q{}^T C_q A$$

$$g_{12} = 0$$

$$g_{13} = -A^T P_1 B - \tau_1 C_q{}^T D_q$$

$$g_{14} = -\tau_2 A^T C_q{}^T D_q$$

$$g_{15} = 0$$

$$g_{22} = P_2 - C_r{}^T Q C_r$$

$$g_{23} = -C_r{}^T (QD_r + S)$$

$$g_{24} = 0$$

$$g_{25} = \Omega_2{}^T A^T (C_2{}^T P_2 - C_1{}^T Y^T)$$

$$g_{33} = \tau_1(I - {D_q}^T D_q) - B^T P_1 B - {D_r}^T Q D_r - {D_r}^T S - S^T D_r - R$$

$$g_{34} = 0$$

$$g_{35} = (B - A\Omega_1 D)^T ({C_2}^T P_2 - {C_1}^T Y^T)$$

$$g_{44} = \tau_2(I - {D_q}^T D_q)$$

$$g_{45} = -D^T Y^T$$

$$g_{55} = P_2$$

for some Y, $P_1 > 0$, $P_2 > 0$, and $\tau_1, \tau_2 \geq 0$, then the state and estimation error dynamics satisfy:

$$\begin{aligned} E\{{x_N}^T P_1 x_N\} + E\{{e_N}^T P_2 e_N\} &\leq E\{{x_0}^T P_1 x_0\} + E\{{e_0}^T P_2 e_0\} \\ &+ \tau_1 \sum_{k=0}^{N-1} E\{\| p_k \|^2 - \| q_k \|^2\} + \tau_2 \sum_{k=0}^{N-1} E\{\| p_{k+1} \|^2 - \| q_{k+1} \|^2\} \\ &- \sum_{k=0}^{N-1} E\{({r_k}^T \quad {p_k}^T) \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \begin{pmatrix} r_k \\ p_k \end{pmatrix} \} \end{aligned} \tag{19}$$

for all $N \in \mathbf{Z}^+$, where the estimator gain is found by $K = {P_2}^{-1} Y$.

Sketch of Proof. Letting $V_k = {x_k}^T P_1 x_k + {e_k}^T P_2 e_k > 0$, consider

$$\begin{aligned} E_{x_k, e_k}\{ V_{k+1} - V_k - \tau_1(\| p_k \|^2 - \| q_k \|^2) - \tau_2(\| p_{k+1} \|^2 - \| q_{k+1} \|^2) \\ + ({r_k}^T \quad {p_k}^T) \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \begin{pmatrix} r_k \\ p_k \end{pmatrix} \} \doteq \Delta V_k \end{aligned} \tag{20}$$

which upon substitution from (15) and (16), taking expectation of both sides, using the interlacing property of expectations and rearranging, leads to

$$E\{\Delta V_k\} \leq -E\{({x_k}^T \quad {e_k}^T \quad {p_k}^T \quad {p_{k+1}}^T) \mathcal{H} \begin{pmatrix} x_k \\ e_k \\ p_k \\ p_{k+1} \end{pmatrix} \} \tag{21}$$

where

$$\mathcal{H} = \begin{pmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ {h_{12}}^T & h_{22} & h_{23} & h_{24} \\ {h_{13}}^T & {h_{23}}^T & h_{33} & h_{34} \\ {h_{14}}^T & {h_{24}}^T & {h_{34}}^T & h_{44} \end{pmatrix}$$

with

$$h_{11} = P_1 - A^T P_1 A - \tau_1 {C_q}^T C_q - \tau_2 A^T {C_q}^T C_q A$$

$$h_{12} = 0$$

$$h_{13} = -A^T P_1 B - \tau_1 C_q{}^T D_q$$

$$h_{14} = -\tau_2 A^T C_q{}^T D_q$$

$$h_{22} = P_2 - A_0{}^T P_2 A_0 - C_r{}^T Q C_r$$

$$h_{23} = -A_0{}^T P_2 B_0 - C_r{}^T (Q D_r + S)$$

$$h_{24} = A_0{}^T P_2 K D$$

$$h_{33} = \tau_1 (I - D_q{}^T D_q) - B^T P_1 B - B_0{}^T P_2 B_0 - D_r{}^T Q D_r - D_r{}^T S - S^T D_r - R$$

$$h_{34} = B_0{}^T P_2 K D$$

$$h_{44} = \tau_2 (I - D_q{}^T D_q) - D^T K^T P_2 K D$$

Since LMI (17) holds, then by Schur's complement result $\mathcal{H} \geq 0$. Therefore,

$$-E\{(x_k{}^T \quad e_k{}^T \quad p_k{}^T \quad p_{k+1}{}^T) \mathcal{H} \begin{pmatrix} x_k \\ e_k \\ p_k \\ p_{k+1} \end{pmatrix} \} \leq 0 \tag{22}$$

Summing over k, leads to (19) which completes the proof.

The above proof of the main result provides a procedure for designing state estimators with improved response. By applying Rayleigh's inequality in (19), one can see that

$$\begin{aligned} \lambda_{min}(P_1) E\{\| x_N \|^2\} + \lambda_{min}(P_2) E\{\| e_N \|^2\} &\leq \lambda_{max}(P_1) E\{\| x_0 \|^2\} + \lambda_{max}(P_2) E\{\| e_0 \|^2\} \\ &+ \tau_1 \sum_{k=0}^{N-1} E\{\| p_k \|^2 - \| q_k \|^2\} + \tau_2 \sum_{k=0}^{N-1} E\{\| p_{k+1} \|^2 - \| q_{k+1} \|^2\} \\ &- \sum_{k=0}^{N-1} E\{(r_k{}^T \quad p_k{}^T) \begin{pmatrix} Q & S \\ S^T & R \end{pmatrix} \begin{pmatrix} r_k \\ p_k \end{pmatrix} \} \end{aligned} \tag{23}$$

So, maximizing $\lambda_{min}(P_2)$, τ_1 and τ_2, minimizing $\lambda_{max}(P_1)$ and $\lambda_{max}(P_2)$ subject to the LMI (17) will give a smaller mean-square error. This is an eigenvalue problem solvable by available LMI software [1].

Application to Various Estimation Problems

Theorem 1 given above allows one to design different estimators for a variety of performance criteria for this class of systems. For example, taking $Q = 0$, $S = 0$, and $R = -\mu I$, $\mu > 0$ yields

$$E\{e_N{}^T P_2 e_N\} \leq E\{x_0{}^T P_1 x_0\} + E\{e_0{}^T P_2 e_0\} + \mu \sum_{k=0}^{N-1} E\{\| p_k \|^2\}$$

This means that by employing the optimization procedure described above, we can obtain a tight bound on the mean-square estimation error.

By taking $Q = \delta I$, $\delta > 0$, $S = 0$, $R = 0$, $B = 0$, $D = 0$, $D_q = 0$ and $D_r = 0$, we obtain

$$\delta \sum_{k=0}^{N-1} E\{r_k{}^T r_k\} \leq E\{x_0{}^T P_1 x_0\} + E\{e_0{}^T P_2 e_0\}$$

which yields a bound on the energy of the performance output in terms of the initial estimation error (suboptimal H_2 result).

If we set $Q = I$, $S = 0$, and $R = -\mu I$, $\mu > 0$, for $x_0 = 0$ and $e_0 = 0$, this produces the result

$$\sum_{k=0}^{N-1} E\{r_k{}^T r_k\} \leq \mu \sum_{k=0}^{N-1} E\{\| p_k \|^2\}$$

which is a bound on the stochastic (mean-square) l_2 to l_2 gain of the estimator (suboptimal H_∞ result).

Several dissipative estimator designs are also possible using this formulation. For example, taking $x_0 = 0$, $e_0 = 0$, $Q = 0$, $S = -0.5I$, and $R = \mu I$, $\mu > 0$ will yield the stochastic (mean-square) version of the input strict passivity result:

$$\sum_{k=0}^{N-1} E\{r_k{}^T p_k\} \geq \mu \sum_{k=0}^{N-1} E\{\| p_k \|^2\}$$

Other similar dissipativity results are also possible. For example, setting $Q = 0$, $S = -0.5I$, and $R = 0$ will give stochastic passivity. Setting $Q = \nu I$, $\nu > 0$, $S = -0.5I$, and $R = 0$ will yield output strict passivity. Also, setting $Q = \nu I$, $S = -0.5I$, $R = \mu I$, ν, $\mu > 0$ will give strict passivity both in terms of the input and the output (very strict passivity in the mean-square sense). So, one can see that this LMI formulation allows one to consider a variety of performance criteria in a common framework.

Conclusion

We have considered reduced-order linear observer (state estimator) designs for a class of discrete-time uncertain stochastic systems with quadratic sum constraints and general performance criteria. We have shown that a common framework using linear matrix inequality formulations can be provided to solve diverse estimator design problems. The future work will involve the robustness study of these estimators in the presence of uncertain parameters and erroneous statistical knowledge of the stochastic processes involved.

References

1. S. Boyd, L. El Ghaoui, E. Feron and V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory, SIAM: Philadelphia, PA (1994).
2. D. H. Jacobson, "A general result in stochastic optimal control of nonlinear discrete-time systems with quadratic performance criteria," J. Math. Anal. Applic., vol. 47, pp. 153 - 161 (1974).

3. Y. I. Yaz and E. E. Yaz, "On LMI formulations of some problems arising in nonlinear stochastic system analysis," IEEE Trans. Autom. Control, vol. 44, pp. 813 - 816 (1999).
4. E. Yaz, "Infinite horizon quadratic optimal control of a class of nonlinear stochastic systems," IEEE Trans. Autom. Control, vol. 34, pp. 1176 - 1180 (1989).
5. E. Yaz and Y. I. Yaz, "Stochastic Stabilizability of a Class of Discrete-Time Nonlinear Stochastic Systems," Proc. of 1999 IFAC World Congress, Beijing, China, vol. J, pp. 253-258 (1999).
6. E. Yaz, "Linear state estimators for nonlinear stochastic systems with noisy nonlinear observations," Int. J. Control, vol. 48, pp. 2465 - 2475 (1988).
7. E. Yaz, "Robust design of stochastic controllers for nonlinear systems," IEEE Trans. Autom. Control, vol. 34, pp. 349 - 353 (1989).
8. A. Ben-Israel and T. N. E. Greville, Generalized Inverses: Theory and Applications, Wiley: New York (1974).

A Simplified Time Maximum Disturbance Switch Curve for Second Order Systems

K. H. You and E. B. Lee
ykh@ece.umn.edu and eblee@ece.umn.edu
Department of Electrical and Computer Engineering
University of Minnesota, 200 Union Street
Minneapolis, MN 55455

Abstract

The most stressful bounded disturbance (time-delay maximizing) for a linear stable second order system with(or without) numerator dynamics is a bang-bang type disturbance and should be realized in state feedback form using a switch curve. In this paper, to overcome some of the complications of the general disturbance switch curve, we have found an approximate disturbance switch line which is almost as good. This simpler switch line requires less computational effort and improves system robustness. Simulation results show the similarity to the original disturbance switch curve when in action and proximate maximization of the time severity index.

Key words: Time maximum disturbance, bang-bang, switch curve, damped harmonic oscillator.

1 Introduction

The switch curve for the time-delay maximizing disturbance for a second order damped linear oscillator consists of an infinite number of sections, smooth curves between kink points. Previously, this overall switch curve has been found by integrating a set of system equations backwards in time. This procedure is complicated and computationally intensive. Recently, a new form of the switch curve was found based on a scaling factor [3]-[6] in closed analytic form. This new method requires much less computation and gives a better intuitive feel for how the curve will change with each parameter. To even further simplify the switch curve an approximation which is almost as stressful when used to realize the bounded disturbance will be given in this paper.

Time optimal control has been an important approach when transferring a system from any arbitrary initial state to a desired goal state with bounded control. From the maximal principle and the method of dynamic programming, the control function has been found in state feedback form(bang-bang type). For stress testing a stabilized system with bounded disturbance, it is also known that the time-delay maximizing disturbance is a bang-bang feedback signal.

Many systems can be adequately modeled using a second order differential equation. Recent effort has found the feedback disturbance in closed analytic form for the second order systems (Lee and Luo, [2], [3]). Also for the second order systems with numerator dynamics in the transfer function, there is a closed analytic form for the disturbance switch curve [5]. The disturbance switch curves(see ref.[3], figure 1, ref.[5], figure 1) are smooth curves between kink points. In the region of this sharp curvature, the system is vulnerable to model uncertainties. This results in a decrease in the time-delay associated with maximizing severity. Here we derive an approximate disturbance switch line from the solutions of second order systems directly which overcomes some of these difficulties. A similar approximation of the switch curve was given for the time optimal control of a family of systems in ref. [4].

It is known that time optimal feedback control for some second order systems is acutely sensitive to parameter perturbations. In fact there arises limit cycle oscillations for realized feedback system when perturbed [7]. Pao has proposed a proximate time-optimal controller for third-order servomechanisms($PTOS$) which are robust to parameter uncertainties and nearly time-optimal [8]. The approximate time maximum feedback disturber found in this paper is quite insensitive to such changes in the system model as well. Kogan has proved the necessary and sufficient conditions for the controllers and disturbances to be solutions of a certain linear quadratic minimax control task [9]. The controller does minimization and the disturbance does maximizing; selected in the game theoretic approach related to H_∞ control. Our results are related to this minimax approach.

2 Approximation of Disturbance Switch Curve

2.1 Harmonic oscillator of second order with damping

Consider the damped harmonic oscillator with output $y(t)$ and bounded input $v(t)$ as given by linear second order input/output differential equation

$$\ddot{y}(t) + 2\zeta\omega_n\dot{y}(t) + \omega_n^2 y(t) = {\omega_n}^2 v(t). \tag{1}$$

Here $0 < \zeta < 1$ is the damping ratio. Without loss of generality, the undamped natural frequency ω_n is taken to be 1. $|v(t)| \leq 1$ is a bounded disturbance signal to stress the stabilized linear second order system as much as possible in the sense of maximum time delay. The most stressful disturbance is a bang-bang state feedback type signal [1] [2] [3]. The dynamical equation with state variables ($x_1(t) \triangleq y(t)$, $x_2(t) \triangleq \dot{y}(t)$) is obtained as

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -x_1(t) - 2\zeta x_2(t) + v(t), \end{aligned} \tag{2}$$

with input $v(t)$ and output $x_1(t)$.

Our task now is: *Find the most stressful disturbance which maximizes the severity index*

$$\mathbf{J} = max \int_0^{t_f} dt.$$

subject to the system equation, equation (2) and various end point constraints. Using the maximum principle (see ref.[1], page 212), we define a Hamiltonian function H by setting

$$H = 1 + \eta_1(x_2) + \eta_2(-x_1 - 2\zeta x_2 + v)$$

where the corresponding adjoint system is

$$\begin{aligned} \dot{\eta_1}(t) &= \eta_2(t) \\ \dot{\eta_2}(t) &= -\eta_1(t) + 2\zeta\eta_2(t). \end{aligned} \tag{3}$$

With the bounded disturbance input, the disturbance maximizing the severity index is in feedback form with a switching based on $\eta_1(t)$ and $\eta_2(t)$ as

$$v(t) = sgn[\eta_2(t)],$$

where $sgn(\cdot)$ means a signum function. Figure 1 shows an example of constructing the disturbance signal with the associated switch curve. The switch curve is constructed backwards in time with the adjoint system equation (see ref. [1], page 434).

Next we consider how to approximate the disturbance switch curves. The solution of the system (equation (2)) as a function of time for constant disturbance ($v(t) \equiv 1$) and initial conditions ($x_1(0) = 0, x_2(0) = 0$) at the origin for one non-switching interval $0 \leq t \leq \frac{\pi}{\sqrt{1-\zeta^2}}$ is

$$\begin{aligned} x_1(t) &= \Delta - \Delta e^{-\zeta t}\cos(\sqrt{1-\zeta^2}t) - \frac{\Delta\zeta}{\sqrt{1-\zeta^2}}e^{-\zeta t}\sin(\sqrt{1-\zeta^2}t) \\ x_2(t) &= \frac{\Delta}{\sqrt{1-\zeta^2}}e^{-\zeta t}\sin(\sqrt{1-\zeta^2}t) \end{aligned} \tag{4}$$

with $\Delta = 1$. When $v(t) = -1$, use $\Delta = -1$. The solution segment equation (4) for $0 \leq t \leq \frac{\pi}{\sqrt{1-\zeta^2}}$, is the closed analytic form of the first part of the disturbance switch curve in parametric form for second order systems (see ref.[3]). Figure 1 shows the disturbance switch curve from the calculation of a closed analytic form showing the first segments and subsequent segments obtained as a simple mapping of logarithmic spiral segments by the logarithmic spiral solution curves [2] [3]. In this figure, we see that the kink points approach to (A_∞, B_∞) with $A_\infty = \frac{1+e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}{1-e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}$ and $B_\infty = 0$ (see ref.[3]).

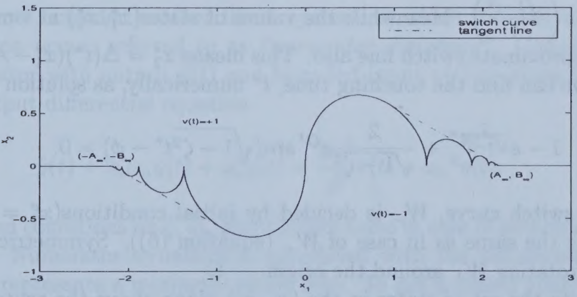


Figure 1: Disturbance switch curve and tangent line from (A_∞, B_∞) with $\zeta = 0.3$.

Especially around the kink points, there exists large changes in the curvature. If the state values come close to this kink region, even with small amount of model uncertainty, it is hard to maximize the performance index. We have developed another disturber which is more robust to model uncertainty by using an approximate switch line as shown in figure 1(the dash-dot line segment). The closed analytic form of the approximation to the disturbance switch curve is given next.

Proposition 1 *The approximate switch line inside of the maximum limit cycle* Γ^* *is* $W(x_1, x_2) = W_+ \cup W_-$, W_+ *means the half of switch curve constructed by* $v(t) = 1$ *as in equation (4). For* $0 \leq t < t^*$, *the switch curve is determined by equation (4) and beyond* t^*, *the approximate disturbance switch line is given as (straight line)*

$$\begin{aligned} x_2 &= \Delta(t^*)[x_1 - A_\infty] + B_\infty \\ where\ \Delta(t^*) &= -\frac{sin(\sqrt{1-\zeta^2}t^* - \phi)}{sin(\sqrt{1-\zeta^2}t^*)} \\ and\ \phi &= tan^{-1}\left(\frac{\sqrt{1-\zeta^2}}{\zeta}\right). \end{aligned} \tag{5}$$

Here t^* *from equation (6) is the time when the disturbance switch line becomes a tangent line to the first disturbance switch curve as determined by equation (4).* $\Delta(t^*)$ *is the slope of the tangent line.* W_- *are the other half of switch curve constructed by* $v(t) = -1$. *Further more,* W_+ *and* W_- *are symmetric through the origin(clarification of this situation in figure 1).*

Proof : In figure 2, there is a curve segment labeled as W_+ inside of maximum limit cycle Γ^*. W_+ is decided by the solution of equation (4) with initial condition($x_1^o = 0, x_2^o = 0$) and $v(t) = 1$ for $0 \leq t \leq t^*$. After the touching time t^*, the switch curve is replaced by the tangent line that meets each switch curve segments. The equation of slope as a function of time can be found from the relation, $\Delta(t) = \frac{dx_2}{dx_1} = \frac{dx_2}{dt} / \frac{dx_1}{dt}$,

$$\begin{aligned} \Delta(t) &= \frac{\sqrt{1-\zeta^2}cos(\sqrt{1-\zeta^2}t) - \zeta sin(\sqrt{1-\zeta^2}t)}{sin(\sqrt{1-\zeta^2}t)} \\ &= -\frac{sin(\sqrt{1-\zeta^2}t - \phi)}{sin(\sqrt{1-\zeta^2}t)} \end{aligned}$$

where $\phi = tan^{-1}\left(\frac{\sqrt{1-\zeta^2}}{\zeta}\right)$. Meanwhile the values of states(x_1^*, x_2^*) at time $t = t^*$ are the points on the approximate switch line also. This means $x_2^* = \Delta(t^*)(x_1^* - A_\infty) + B_\infty$. With this equation, we can find the touching time, t^* numerically, as solution of equation

$$1 - e^{\frac{\zeta}{\sqrt{1-\zeta^2}}\pi} + \frac{2}{\sqrt{1-\zeta^2}} e^{\zeta t^*} sin(\sqrt{1-\zeta^2}t^* - \phi) = 0. \tag{6}$$

The other half switch curve, W_- is decided by initial conditions($x_1^o = 0, x_2^o = 0$) and $v(t) = -1$. t^* is the same as in case of W_+ (equation (6)). Symmetrically W_- can be constructed by rotating W_+ around the origin. □

Let R_+ denote the set of states in the (x_1, x_2) plane above the switch curve W and let R_- denote the set of states below the switch curve W as shown in Figure 2. That is,

$$R_+ = \{(x_1, x_2) : if\ (x_1, \hat{x}_2) \in W,\ then\ x_2 > \hat{x}_2\}$$

$$R_- = \{(x_1, x_2) : if\ (x_1, \hat{x}_2) \in W,\ then\ x_2 < \hat{x}_2\}.$$

We assign $v^* = +1$ on W_+ and $v^* = -1$ on W_-. So the maximal time nastiest disturbance as a function of the state is given by

$$v^* = \begin{cases} v^*(x_1, x_2) = +1, & \text{for all } (x_1, x_2) \in R_+ \cup W_+ \\ v^*(x_1, x_2) = -1, & \text{for all } (x_1, x_2) \in R_- \cup W_-. \end{cases} \tag{7}$$

Figure 2 shows the approximate disturbance switch line. In this figure, the outer closed curve that encloses the switch line is a maximum limit cycle (see ref.[3]). The switch curve outside of the maximum limit cycle Γ^* is given as x_1-axis (*i.e.* x_2=0) just as before [3] and need not be approximated. With this disturbance switch curve(line), any initial point around the kink region will give a trajectory approaching to the maximum limit cycle and oscillates along this cycle with switching time, $T_s = \frac{\pi}{\sqrt{1-\zeta^2}}$. Simulation results involving this approximate limit cycle will be given in section 3. We now generalize the results to the case with a finite zero.

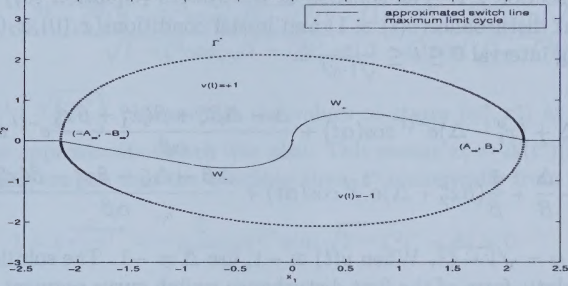


Figure 2: Approximate disturbance switch line with $\zeta = 0.3$.

2.2 Harmonic oscillator of second order system with damping and numerator dynamics

In this section, we consider another linear system, in which the derivatives of the disturbance function occur, referred to as *"numerator dynamics"*. Consider the damped harmonic oscillator with output $y(t)$ and bounded input $v(t)$ as given by linear second order input/output differential equation

$$\ddot{y}(t) + 2\zeta\omega_n\dot{y}(t) + \omega_n^2 y(t) = \frac{{\omega_n}^2}{\beta}\dot{v}(t) + {\omega_n}^2 v(t). \tag{8}$$

The notation and conditions on ζ, ω_n and $v(t)$ are same in case of no numerator dynamics in section 2.1. Numerator dynamics is associated with the parameter $\beta < \infty$. For $\beta > 0$, $s = -\beta$ represents a minimum phase zero. To get a closed analytic form for the disturbance switch curve, we now need suitable state variables without jumps; so take $x_1(t) \triangleq y(t)$, $x_2(t) \triangleq \dot{y}(t) - \frac{1}{\beta}v(t)$ (see ref.[1], page 35). The input/output system with new state variables is

$$\dot{x_1}(t) \;=\; x_2(t) + \frac{1}{\beta}v(t)$$

$$\dot{x}_2(t) = -x_1(t) - 2\zeta x_2(t) + (1 - 2\frac{\zeta}{\beta})v(t), \quad (9)$$

with input $v(t)$ and output $x_1(t)$. Following the steps in section 2.1, the Hamiltonian with the same adjoint system equation (3) is

$$H = 1 + \eta_1(x_2 + \frac{1}{\beta}v) + \eta_2[-x_1 - 2\zeta x_2 + (1 - 2\frac{\zeta}{\beta})v].$$

From maximal principle, the maximal time disturbance is of the form

$$v(t) = sgn[\frac{1}{\beta}\eta_1(t) + (1 - 2\frac{\zeta}{\beta})\eta_2(t)].$$

Figure 3 shows an example of disturbance switch curve constructed backwards in time as explained in section 2.1. The solution of the system (equation (8)) as a function of time for constant disturbance($v(t) \equiv 1$) and initial conditions($x_1(0), x_2(0)$) as the origin for one switching interval $0 \leq t \leq \frac{\pi}{\sqrt{1-\zeta^2}}$ is

$$\begin{aligned} x_1(t) &= \Delta + (x_1^o - \Delta)e^{-\zeta t}\cos(\alpha t) + \frac{\Delta - \Delta\beta\zeta + \beta\zeta x_1^o + \beta x_2^o}{\alpha\beta}e^{-\zeta t}\sin(\alpha t) \\ x_2(t) &= -\frac{\Delta}{\beta} + \frac{1}{\beta}(\beta x_2^o + \Delta)e^{-\zeta t}\cos(\alpha t) + \frac{\Delta\beta - \Delta\zeta - \beta x_1^o - \beta\zeta x_2^o}{\alpha\beta}e^{-\zeta t}\sin(\alpha t) \quad (10) \end{aligned}$$

with $\Delta = 1$ and $\alpha = \sqrt{1-\zeta^2}$. When $v(t) \equiv -1$, use $\Delta = -1$. The solution equation (10) is the closed analytic form of the first disturbance switch curve segment for second order system (see ref.[5]). Figure 3 shows the disturbance switch curve from the calculation of a closed analytic form. In this figure, we see that the kink points approach to (A_∞, B_∞) with $A_\infty = \frac{1+e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}{1-e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}$ and $B_\infty = \frac{-1-e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}{\beta-\beta e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}$ (see ref.[5]). Now we seek a closed

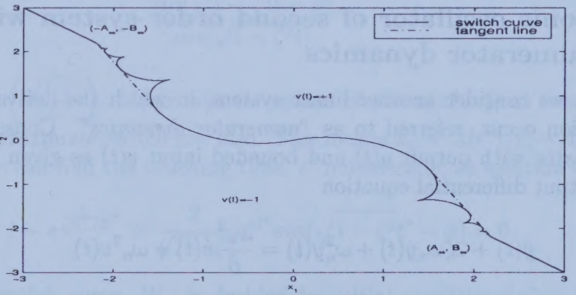


Figure 3: Disturbance switch curve and tangent line from (A_∞, B_∞) with $\beta = 1, \zeta = 0.3$.

analytic form for the straight line segment approximation to the disturbance switch curve shown as dash-dot curve in figure 3.

Proposition 2 *Consider W, W_+ and W_- to be the same as in proposition 1. With numerator dynamics, $0 < \beta < \infty$, the equation of approximate disturbance switch line is* *changed to*

$$x_2 = \Delta(t^*)[x_1 - A_\infty] + B_\infty$$

Where

$$\Delta(t^*) = \frac{\sqrt{1-\zeta^2}(\beta - 2\zeta)cos(\sqrt{1-\zeta^2}t^*) - (1+\zeta\beta - 2\zeta^2)sin(\sqrt{1-\zeta^2}t^*)}{\sqrt{1-\zeta^2}cos(\sqrt{1-\zeta^2}t^*) + (\beta - \zeta)sin(\sqrt{1-\zeta^2}t^*)}. \quad (11)$$

Here t^* is the time when the tangent line meets with the first disturbance switch curve as determined by equation (12). $\Delta(t^*)$ is the slope of the tangent line. Also, W_- can be found through rotating W_+ around the origin (rotate by 180^o).

Proof : The function of slope in time can be found from the relation, $\Delta(t) = \frac{dx_2}{dx_1} = \frac{dx_2}{dt} / \frac{dx_1}{dt}$,

$$\Delta(t) = \frac{\sqrt{1-\zeta^2}(\beta - 2\zeta)cos(\sqrt{1-\zeta^2}t) - (1+\zeta\beta - 2\zeta^2)sin(\sqrt{1-\zeta^2}t)}{\sqrt{1-\zeta^2}cos(\sqrt{1-\zeta^2}t) + (\beta - \zeta)sin(\sqrt{1-\zeta^2}t)}$$

where $\phi = tan^{-1}\left(\frac{\sqrt{1-\zeta^2}}{\zeta}\right)$. Meanwhile the values of states (x_1^*, x_2^*) at time $t = t^*$ are the points on the approximate switch line also. This means $x_2^* = \Delta(t^*)(x_1^* - A_\infty) + B_\infty$. With this equation, we can find the touching time, t^* numerically from the equation.

$$1 - e^{\frac{\zeta}{\sqrt{1-\zeta^2}}\pi} + \frac{2}{\sqrt{1-\zeta^2}} e^{\zeta t^*} sin(\sqrt{1-\zeta^2}t^* - \phi) = 0 \quad (12)$$

This completes the proof of proposition 2. □

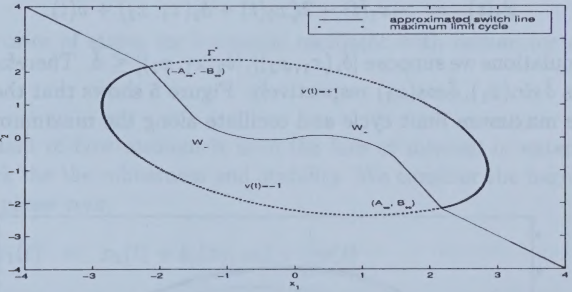


Figure 4: Approximate Disturbance switch line with $\beta = 1$, $\zeta = 0.3$.

Figure 4 shows the approximate disturbance switch line. In this figure, the outer closed curve that encloses the switch line is a maximum limit cycle (see ref.[5]). With this disturbance switch curve (line), any initial points approach to the maximum limit cycle and oscillates along this cycle with switching time, $T_s = \frac{\pi}{\sqrt{1-\zeta^2}}$. The disturbance switch curve for maximum time-delay outside of maximum limit cycle Γ^* is again a straight line [5] (given by $x_2 = -\frac{1}{\beta}x_1$) and need not be approximated.

Remark: If $\beta \to \infty$ in equation (8), control system becomes a harmonic oscillator without numerator dynamics. The switch construction process shown in Proposition 2 is in accord with that in Proposition 1. Moreover, the BIBO stability integral (L_∞-gain) turns to the distance between origin and the final kink point $A_\infty = \frac{1+e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}{1-e^{-\frac{\zeta}{\sqrt{1-\zeta^2}}\pi}}$, $B_\infty = 0$ as $\beta \to \infty$ (see ref. [5]).

3 Simulation Results

The results obtained above can be used to construct the approximate disturbance switch line with only one curved part as shown in figure 2, 4. Through the following two cases, we now show how useful the approximate switch line is as a substitute for the original switch curve.

Case 1: We simulate the harmonic oscillator of a second order system without numerator dynamics (at $\zeta = 0.3$). With the closed analytic form for the approximate disturbance switch line in figure 2, all state responses approach the maximum limit cycle as they should. Some uncertainties(parameter uncertainty, modeling error, external noise) can diminish the system robustness and make it hard to test exactly. For this reason, the approximate disturbance switch line gives added benefit. First, it is easy to construct the disturbance switch line. It saves computational time with fewer segmentation decision curves involved. Secondly without kink changes in curvature in disturbance switch line, it is less likely to have switching errors even with some uncertainties.

We show the robustness of the proposed disturbance generating procedure for a given system with model uncertainty. Now we suppose that the model uncertainty is added to the dynamic system, equation as :

$$\begin{aligned}\dot{x_1}(t) &= x_2(t) + \delta_1(x_1, x_2) \\ \dot{x_2}(t) &= -x_1(t) - 2\zeta x_2(t) + \delta_2(x_1, x_2) + v(t).\end{aligned} \tag{13}$$

Through the simulations we suppose $|\delta_1(x_1, x_2)|, |\delta_2(x_1, x_2)| \leq \bar{\delta}$. Therefore $\delta_1(\cdot), \delta_2(\cdot)$ can be thought of as $\bar{\delta} sin(x_1), \bar{\delta} cos(x_2)$ respectively. Figure 5 shows that the state responses approach to the maximum limit cycle and oscillate along the maximum limit cycle as t increases.

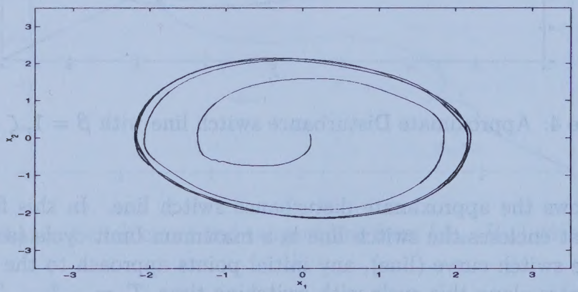


Figure 5: Trajectories of states starting from inside of maximum limit cycle with model uncertainty, $\zeta = 0.3, \bar{\delta} = 0.1$.

Case 2: Next consider the harmonic oscillator simulation of a second order system with numerator dynamics($\beta = 1$, $\zeta = 0.3$). An approximate disturbance switch line was constructed following the steps in proposition 2.

Figure 6 shows the effectiveness of the approximate disturbance switch line. As established in case of no numerator dynamics, all states converge to the maximum limit cycle and move as shown in figure 6. Again the state trajectories determined by the new approximate switch line satisfy the time severity index approximately.

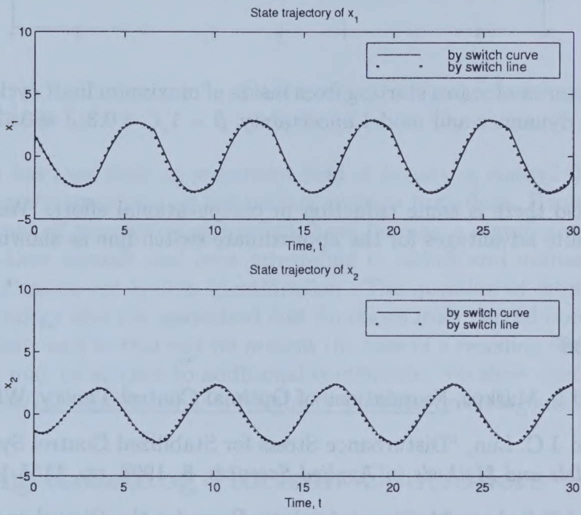


Figure 6: Trajectories of states for harmonic oscillator with numerator dynamics, $\beta = 1, \zeta = 0.3$.

As an assessment of how resilient it is in the face of internal or external changes in behavior, we check the the robustness and stability. We consider the harmonic oscillator with a minimum phase zero,

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) + \delta_1(x_1, x_2) + \frac{1}{\beta}v(t) \\ \dot{x}_2(t) &= -x_1(t) - 2\zeta x_2(t) + \delta_2(x_1, x_2) + (1 - 2\frac{\zeta}{\beta})v(t) \end{aligned} \qquad (14)$$

$\delta_1(\cdot), \delta_2(\cdot)$ are the model uncertainties and the upper bounds are supposed to be $\bar{\delta}$ as in Case 1. Figure 7 shows that the state trajectory comes close to the maximum limit cycle when we use the approximate disturbance switch line for the system with bounded model uncertainty.

4 Conclusions

In this paper, we have found the closed analytic form of an approximate disturbance switch line for the maximum time-delay severity index. With the new proposed approximation of the disturbance switch curve, an improvement in the system robustness is

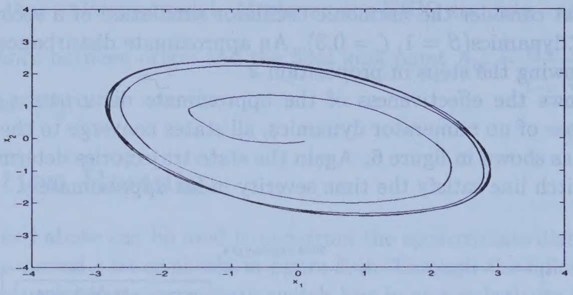


Figure 7: Trajectories of states starting from inside of maximum limit cycle for the system with numerator dynamics and model uncertainty, $\beta = 1, \zeta = 0.3, \bar{\delta} = 0.1$.

obtained and also there is some reduction in computational effort. We have therefore established definite advantages for the approximate switch line as shown by the simulations.

References

[1] E.B. Lee and L. Markus, Foundations of Optimal Control Theory, Wiley, N. Y., 1967

[2] E.B. Lee and J.C. Luo, "Disturbance Stress for Stabilized Control Systems," *Mathematical Models and Methods in Applied Sciences,* **8**, 1998, pp. 1155-1182.

[3] J.C. Luo and E.B. Lee, "A Closed Analytic Form for the Disturbance switch Curve of Some Second Order Linear Systems," *IEEE Trans. on Automat. Contr.*, **45**, 2000, pp. 102-105.

[4] G. Gu and E.B. Lee, "Design of a Common Feedback Controller for a Family of Nonlinear Second Order Systems," *23rd Annual Allerton Conference*, 1985, pp. 113-122.

[5] K.H. You and E.B. Lee, "BIBO Stability Integral (L_∞-gain) for Second Order Systems with Numerator Dynamics," to appear on *Automatica*, Nov., 2000.

[6] K.H You and E.B. Lee, "Time Maximum Disturbance Switch Curve and Isochrones of Second Order Linear Systems with Numerator Dynamics," to appear on *J. Franklin Inst.*

[7] E.P. Ryan, "On the Sensitivity of a Time-Optimal Switching Function," *IEEE Trans. Automat. Contr.*, **25**, 1980, pp. 275-277.

[8] L.Y. Pao and G.F. Franklin, "The Robustness of a Proximate Time-Optimal Controller," *IEEE Trans. Automat. Contr.,* **39**, 1994, pp. 1963-1966.

[9] M. Kogan, "Solution to the Inverse Problem of Minimax Control and Worst Case Disturbance for Linear Continuous-Time Systems," *IEEE Trans. Automat. Contr.*, **43**, 1998, pp. 670-674.

Identifying Constrained Receding Horizon Controllers *

Sommer Gentry Saligrama Venkatesh Eric Feron
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA 02139

1 Introduction

System identification has long been an important field of inquiry in control theory [4] and has primarily focused on identifying dynamical behavior of systems from data. Recently, receding horizon control and other types of procedural controllers have become a significant element in chemical engineering and real-time aircraft and crew scheduling in airline and military operations [1] and this offers a new application for system identification. The question of interest is to identify the optimizing control strategy and the associated cost functions from partial observations of an "optimized" dynamical system and to this end we present the case of a receding horizon controller where the control and state may be subject to additional constraints. We show that under certain conditions, this problem may be formulated as a feasibility problem involving semidefinite constraints.

2 Identifying Receding Horizon Controllers

Consider the system $\{x(k+1)=Ax(k)+Bu(k),\ Cx(k)+Du(k)\leq b,\ \forall k=0,\ldots,\}$ where $x(k)\in\mathcal{R}^n$ is the state, $u(k)\in\mathcal{R}^p$ is the control input. The matrices $C\in\mathcal{R}^{m\times n}$, $D\in\mathcal{R}^{m\times p}$ and $B\in\mathcal{R}^{m\times 1}$ model linear constraints that may apply to the system. These constraints may include control amplitude and rate constraints as well as forbidden state regions. (The inequality sign appearing above is to be interpreted row-wise.)

A full-state feedback receding horizon controller with horizon length N is assumed to control the system: At every time k, given the state $x(k)$, the input $u(k)$ is selected by solving the optimization problem:

$$\begin{array}{rrcl} \text{Minimize} & \multicolumn{3}{l}{\sum_{i=1}^{N}\hat{x}(i)^T Q\hat{x}(i)+\hat{u}\ (i)^T R\hat{u}(i)} \\ & \hat{x}(1) & = & x(k) \\ \text{subject to} & \hat{x}(i+1) & = & A\hat{x}(i)+B\hat{u}(i),\ \ i=1,\ldots,N \\ & C\hat{x}(i)+D\hat{u}(i) & \leq & b,\ \ i=1,\ldots,N. \end{array} \tag{1}$$

Define the decision variable $\hat{u}_k=[\hat{u}(1),\ldots,\hat{u}(N)]^T$. Let $\hat{u}_k^*$ be the optimal solution to (1); then the control applied to the system at time k is $u(k)=\hat{u}_k^*(1)$.

Problem Formulation: The problem is formulated based on the so called Information Based Complexity (see [3]) approach: Given observations, $\{x_k,\ u_k\}_{k=0}^{n}$, along with the dynamical equations and state and control constraints, find the set of all matrices, Q, R, which are consistent with observations. The formulation can be generalized in many directions. We can deal with noisy observations, worst-case or probabilistic settings and formulate conditions on the input (or an exogenous input) to obtain identifiability. We omit these generalizations here for the sake of brevity.

---

*corresponding Author: Professor Eric Feron, email:feron@mit.edu

We introduce the vector $\hat{x}_k = [\hat{x}(1), \ldots, \hat{x}(N)]^T$ for each time step, k, and aggregate the dynamics as a static map, i.e.: $\hat{x}_k = \mathcal{A}x(k) + \mathcal{B}\hat{u}_k$. The resulting quadratic optimization problem can be rewritten as:

$$(\mathcal{A}x(k) + \mathcal{B}\hat{u}_k)^T \mathcal{Q}(\mathcal{A}x(k) + \mathcal{B}\hat{u}_k) + \hat{u}_k^T \mathcal{R}\hat{u}_k$$

subject to

$$\mathcal{C}(\mathcal{A}x(k) + \mathcal{B}\hat{u}_k) + \mathcal{D}\hat{u}_k \leq \beta$$

where the decision variable is now $\hat{u}_k$ with the appropriate definition for $\mathcal{Q}$, $\mathcal{R}$, $\mathcal{C}$, $\mathcal{D}$, β. Recall that the problem is solved at each time step and new information gets added because $x(k)$ evolves with time leading to new indirect information about the matrices Q and R.

The Kuhn-Tucker conditions for the optimization problem presented above may be written as

$$\begin{aligned} \mathcal{B}^T \mathcal{Q}(\mathcal{A}x(k) + \mathcal{B}\hat{u}_k) + \mathcal{R}\hat{u}_k + (\mathcal{D} + \mathcal{C}\mathcal{B})^T \lambda &= 0 && (2) \\ \lambda^T(\mathcal{C}(\mathcal{A}x(k) + \mathcal{B}\hat{u}_k) + \mathcal{D}\hat{u}_k - \beta) &= 0 && (3) \end{aligned}$$

where $\lambda \in \mathcal{R}^{Nm} \geq 0$ is the Lagrange multiplier. Under the restrictive assumption of known control and state vector for the entire horizon length, the Kuhn-Tucker conditions given above are linear in the parameters Q, R and λ (all other parameters are presumed known). The conjunction of these sets of conditions forms a set of linear constraints that Q and R must satisfy, in addition to the semidefinite constraints $Q \geq 0$ and $R > 0$. Thus, finding the set of all matrices Q and R that do not contradict the observations is a feasibility problem involving semidefinite constraints [2].

3 Conclusion

This paper is concerned with identifying the optimizing strategies and associated costs of optimal control systems such as receding horizon controllers. While we have presented an initial formulation of this problem, we will explore several directions including questions of informativity of data, dealing with noisy observations and concept of persistent excitation for identifying optimizing strategies.

References

[1] Andersson, Kari. *Potential Benefits of Information Sharing During the Arrival Process at Hub Airports* Masters thesis, Massachusetts Institute of Technology, 2000.

[2] Boyd, S.; El Ghaoui, L.; Feron, E.; and Balakrishnan, V. *Linear Matrix Inequalities in System and Control Theory* Philadelphia : Society for Industrial and Applied Mathematics, 1994.

[3] Traub, J. F., Wasilkowski, G. W., and Wozniakowski, H., *Information Based Complexity: Computer Science and Scientific Computing* Academic Press, INC, San Diego 1988.

[4] Ljung, Lennart. *System Identification: Theory for the User* Englewood Cliffs, N.J. : Prentice Hall, 1987.

Stochastic Control under Communication Constraints

Sekhar Tatikonda and Sanjoy Mitter

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We formulate a general distributed control problem. We discuss the differences between signal knowledge, system knowledge, and objective knowledge. We also discuss the differences between centralized and decentralized design. For the case of centralized design we show that the distributed control problem can be solved by using dynamic programming.

1 Introduction

We provide a general framework for modeling distributed control problems. The framework we present allows us to treat plants and communication channels on an equal footing. Similarly we can treat controllers, channel encoders, and channel decoders on an equal footing. As a result we are able to present a conceptual view of the design of a large class of distributed control problems where sensors, controller, and actuators are interconnected through communication channels. See figure 1.

In general a model of a system consists of a set of stochastic kernels, an information pattern, and causality constraints codified by a time-ordering on the variables of interest. These stochastic kernels can represent plant transitions or channel transmissions. These different kernels provide a partial specification of a joint measure over all the random variables of interest. The job of a designer then is to "complete" this joint measure by providing appropriate interconnections between the given plants and channels. Note that this completion must respect the information pattern and causality constraints imposed by the model. In practice these interconnections will be provided through the design of a controller, encoder, or decoder.

The design of the optimal controller for the distributed system can be done in many ways. We distinguish between centralized design and decentralized design. Centralized design occurs when the design of a given controller is done with full knowledge of the other controllers. In decentralized design we do not assume this knowledge. We will show that centrally designed systems can be solved via dynamic programming.

In section 2 we provide our general framework. In section 3 we show that centrally designed control problems can be formulated as a dynamic programming problems. We conclude in section 4.

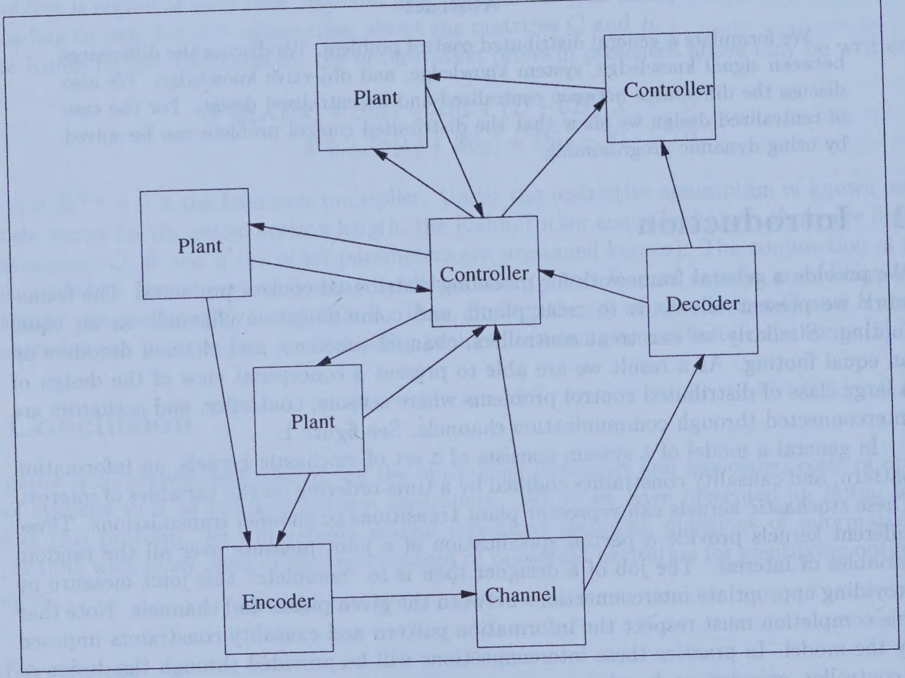

Plant
Controller
Plant
Controller
Decoder
Plant
Encoder
Channel

Figure 1: Distributed System

2 General Framework

We now give our general framework. We assume, through some sort of modeling process, that we are given partial knowledge of the system variables. This partial knowledge comes in the form of a set of stochastic kernels on the variables of interest. From this partial knowledge we want to complete the joint measure on these variables. There are many ways to complete this measure. We define a *model* of a system to be the set of all joint measures consistent with the given stochastic kernels. Thus a model is a subset of the set of all measures on the variables of interest. We now give the formal definition of a model.

Given a sequence of variables $x_1, ..., x_T$ we use the notation $x^T \triangleq (x_1, ..., x_T)$.

2.1 Main Definition

Let $(\Omega, \mathcal{F})$ be a measure space. For $t = 1, ..., T$ let Z_t be a measurable function from $(\Omega, \mathcal{F})$ to the Polish space $(\mathcal{Z}_t, \mathcal{B}(\mathcal{Z}_t))$. (Note that T may equal infinity.) These variables, $\{Z_t\}$, represent different objects in the model. An object, for example, may be the state of a system at a given time, the control at a given time, the channel input at a given time, etcetera. We call them "variables" as opposed to "random variables" because we have not defined a measure on them. Specifically the Z_t's will be random variables only after we define a measure on $(\Omega, \mathcal{F})$. We now define a *model*:

Definition 2.1 *A* model, $\mathcal{M}$, *is a subset of the set of all measures on the variables* $Z_1, ..., Z_T$. *This subset,* $\mathcal{M} \subset \mathcal{P}(Z_1, ..., Z_T)$, *is defined by a three-tuple: causal ordering, system specification, and information pattern.*

(1) Causal Ordering: This is a specification of a time-ordering on the variables $Z_1, ..., Z_T$.

Note that any measure $P(dZ^T)$ *can be factored with respect to this ordering:*

$$P(dZ^T) = \bigotimes_{t=1}^{T} P(dZ_t \mid z^{t-1}).$$

We need to specify these factors. There are two kinds of factors $P(dZ_t \mid z^{t-1})$*. Those that are specified by the system and those that are specified by the designer. (For example state versus control.) Separate the set* $\{1, ..., T\}$ *into these two disjoint sets:*

(a) The system set: $I = \{i_1, ..., i_K\}$ *with* $1 \leq i_1 <, ..., < i_K \leq T$

(b) The decision set: $J = \{j_1, ..., j_L\} = \{1, ..., T\} - I$ *with* $1 \leq j_1 <, ..., < j_L \leq T$.

(2) System specification: This is a set of stochastic kernels of the form $\{Q(dZ_{i_k} \mid z^{i_k-1})\}_{k=1}^{K}$. *For all* $i_k \in I$ *the measure* $P \in \mathcal{M}$ *must satisfy*

$$P(dZ_{i_k} \mid Z^{i_k-1} = z^{i_k-1}) = Q(dZ_{i_k} \mid z^{i_k-1}) \quad P(dZ^{i_k-1}) - a.s.$$

(Where $P(dZ_{i_k} \mid Z^{i_k-1})$ *is the conditional probability under* P.*)*

(3) Information pattern: The information pattern defines what each decision maker $j_l \in J$ *can base his decision on. Specifically for each* $j_l \in J$ *the* information pattern *of decision maker* j_l *is a subset* $J_{j_l} \subseteq \{1, ..., j_l - 1\}$*. The decision kernel* $Q(dZ_{j_l} \mid z^{j_l-1})$ *can only be a function of* $\{z_t : t \in J_{j_l}\}$*. In other words,*

$$Q(dZ_{j_l} \mid z^{j_l-1}) = Q(dZ_{j_l} \mid \tilde{z}^{j_l-1}) \quad \forall z^{j_l-1}, \tilde{z}^{j_l-1} \text{ such that } z_t = \tilde{z}_t \ \forall t \in J_{j_l}.$$

For all $j_l \in J$ the measure $P \in \mathcal{M}$ must satisfy

$$P(dZ_{j_l} \mid Z^{j_l-1} = z^{j_l-1}) = Q(dZ_{j_l} \mid z^{j_l-1}) \quad P(dZ^{j_l-1}) - a.s.$$

The stochastic kernels chosen in part (3) allow us to interconnect the stochastic kernels given in part (2). This interconnection specifies a joint measure. The stochastic kernels chosen in part (3) are called *policies.* Note that an element in $\mathcal{M}$ is uniquely specified, almost surely, once we specify the policies.

Our definition of a model closely parallels Witsenhausen's model. [Wit] In our model, though, we emphasize the two different kinds of stochastic kernels that make up the joint measure: those specified by the system and those that are policies.

Lemma 2.1 *Under any measure $P \in \mathcal{M}$ and $\forall j_l \in J$ the following forms a Markov chain:*

$$Z_{j_l} - \{Z_t \,:\, t \in J_{j_l}\} - \{Z_t \,:\, t \in \{1, .., j_l - 1\} - J_{j_l}.\}$$

Proof: Note that $P(dZ^{j_l}) = P(dZ_{j_l} \mid z^{j_l-1}) \otimes P(dZ^{j_l-1})$. By the information pattern constraint we know $P(dZ_{j_l} \mid Z^{j_l-1} = z^{j_l-1}) = Q(dZ_{j_l} \mid z^{j_l-1}) \quad P(dZ^{j_l-1}) - a.s.$ Thus $P(dZ^{j_l}) = P(dZ_{j_l} \mid \{z_t \,:\, t \in J_{j_l}\}) \otimes P(dZ^{j_l-1})$. But this is the definition of Markov chain. □

The following example will elucidate the previous definition. It is an abstract description of the models used in partially observed Markov decision problems.

Example 2.1 *The variables of interest are $X_1, ..., X_T$, $Y_1, ..., Y_T$ and $U_1, ..., U_T$. The X's are state variables, the Y's are observation variables, and the U's are control variables. The causal ordering is*

$$X_1, Y_1, U_1, ..., X_T, Y_T, U_T.$$

The system specification is given by

$$\{Q(dX_t|x^{t-1}, y^{t-1}, u^{t-1})\}_{t=1}^T \text{ and } \{Q(dY_t|x^t, y^{t-1}, u^{t-1})\}_{t=1}^T.$$

Where $X_t - (X_{t-1}, U_{t-1}) - (X^{t-2}, Y^{t-1}, U^{t-2})$ and $Y_t - X_t - (X^{t-1}, Y^{t-1}, U^{t-1})$ are Markov chains. These are descriptions of the plant and the observation mechanism.

The information pattern specifies $U_t - (Y^t, U^{t-1}) - X^t$ to be a Markov chain. That is the control U_t is allowed to observe only the past observations and controls and not the past states. A policy, then, is a sequence of stochastic kernels $\{Q(dU_t|x^t, y^t, u^{t-1})\}_{t=1}^T$ such that

$$Q(dU_t|x^t, y^t, u^{t-1}) = Q(dU_t|\tilde{x}^t, y^t, u^{t-1}) \quad \forall x^t, \tilde{x}^t.$$

2.2 The Finite Horizon Control Problem

So far we have defined a model. The model, though, only tells us what policies are acceptable. It does not tell us how to choose a policy. To do that we need an objective. Here we define the control objective in terms of minimizing an objective function.

Definition 2.2 *A cost is an integrable function $f : \prod_{t=1}^{T} \mathcal{Z}_t \to I\!R^+$. The control problem consists of computing the cost*

$$\inf_{P \in \mathcal{M}} E_P\left(f(Z_1, ..., Z_T)\right) \tag{1}$$

and finding the infimizing P.

Example 2.1 continued *For the partially observed Markov decision problem the cost is usually of the form*

$$f(X^T, Y^T, U^T) = \sum_{t=1}^{T} c(X_t, U_t)$$

for some "running cost" c.

We are left now with the question of how to compute the optimal solution to the control problem. The solution depends on whether we allow centralized or decentralized design. Roughly speaking centralized design occurs if decision maker j_l can decide $Q(dZ_{j_l} \mid z^{j_l - 1})$ based on the policies of all the other decision makers and complete knowledge of the system specifications. In decentralized design decision maker j_l does not have access to all of the other decision maker's policies or all of the system specifications.

In the next subsection we discuss system and policy knowledge. Then we define centralized and decentralized design.

2.3 System and Policy Knowledge

One component of our model is the information pattern of each decision maker. This information pattern specifies who knows what and when. But that "what" refers to knowledge of the actual signals in the system and not to the knowledge of the different stochastic kernels that make up the joint measure. In computing the solution to (1) we need to determine whether the design of the system, i.e. the specification of the decision kernels, can be done in a centralized manner or a decentralized manner. To get a handle on this we first define what we mean by system and policy knowledge:

Definition 2.3 *The* system knowledge, *$\mathcal{K}_j$, of decision maker j is a subset of*

$$\mathcal{K}_j \subseteq \{Q(dZ_{i_k} \mid z^{i_k - 1}),\ k = 1, ..., K\}.$$

The policy knowledge, *$\mathcal{L}_j$, of decision maker j is a subset of the policies*

$$\mathcal{L}_j \subseteq \{Q(dZ_{j_l} \mid z^{j_l - 1}),\ l = 1, ..., L\}.$$

We further insist that decision maker j knows its own policy: $Q(dZ_j \mid z^{j-1}) \in \mathcal{L}_j$.

Definition 2.4 *We say decision maker j has* complete system knowledge *if*

$$\mathcal{K}_j = \{Q(dZ_{i_k} \mid z^{i_k - 1}),\ k = 1, ..., K\}.$$

And we say decision maker j has complete policy knowledge *if*

$$\mathcal{L}_j = \{Q(dZ_{j_l} \mid z^{j_l - 1}),\ l = 1, ..., L\}.$$

Definition 2.5 *We say a system is* centrally designed *if every decision maker $j_l \in J$ has complete system knowledge and complete policy knowledge. Otherwise call the system* decentrally designed.

In section 3 we will show that centrally designed control problems can be formulated as dynamic programming problems.

2.4 Partial Orderings on the Signal and System Knowledge

We have defined the information pattern and the system and policy knowledge. We now show that there exists a natural partial ordering on the information patterns and the system and policy knowledge.

Recall that an information pattern is defined by the set of sets: $\{J_{j_l} \ : \ l = 1, ..., L\}$. Denote this by $\mathcal{J} \triangleq \{J_{j_l} \ : \ l = 1, ..., L\}$.

Definition 2.6 *We say that $\mathcal{J} \preceq_I \tilde{\mathcal{J}}$ if*

$$J_{j_l} \subseteq \tilde{J}_{j_l} \quad \forall\, l = 1, ..., L.$$

(The "I" in $\preceq_I$ stands for "information pattern.")

It is straightforward to show that $\preceq_I$ is a partial order:

(1) $\mathcal{J} \preceq_I \mathcal{J}$

(2) $\mathcal{J} \preceq_I \tilde{\mathcal{J}}$ and $\tilde{\mathcal{J}} \preceq_I \mathcal{J}$ imply $\mathcal{J} = \tilde{\mathcal{J}}$

(3) $\mathcal{J} \preceq_I \tilde{\mathcal{J}}$ and $\tilde{\mathcal{J}} \preceq_I \hat{\mathcal{J}}$ imply $\mathcal{J} \preceq_I \hat{\mathcal{J}}$

Furthermore there is a a unique maximal element in this partial ordering. Specifically $\mathcal{J}^* = \{J^*_{j_l} \ : \ l = 1, ..., L\ \}$. Where each $J^*_{j_l} = \{1, ..., j_l - 1\}$.

It should be clear that when computing an optimal policy more information cannot increase the optimal cost (recall that an "optimal policy" is a policy that minimizes a given cost.) If $\mathcal{J} \preceq_I \tilde{\mathcal{J}}$ then the optimal cost associated with information pattern $\mathcal{J}$ can be no less than the optimal cost associated with information pattern $\tilde{\mathcal{J}}$. Thus an optimal policy for the maximal information pattern will lead to the minimal cost. It may be the case, though, that there is a smaller, in the sense of $\preceq_I$, information pattern that leads to the same cost. This is an interesting, though, difficult, question.

We can also define a partial ordering on the system and policy knowledge. Recall that the system knowledge and policy knowledge are defined by the two sets of sets: $\{\mathcal{K}_{j_l} \ : \ l = 1, ..., L\}$ and $\{\mathcal{L}_{j_l} \ : \ l = 1, ..., L\}$ respectively. Define $\mathcal{S} \triangleq \{\mathcal{K}_{j_l},\ \mathcal{L}_{j_l} \ : \ l = 1, ..., L\}$.

Definition 2.7 *We say that $\mathcal{S} \preceq_K \tilde{\mathcal{S}}$ if*

$$\mathcal{K}_{j_l} \subseteq \tilde{\mathcal{K}}_{j_l} \ \ and\ \mathcal{L}_{j_l} \subseteq \tilde{\mathcal{L}}_{j_l} \quad \forall\, l = 1, ..., L.$$

(The "K" in $\preceq_K$ stands for "system and policy knowledge.")

It is straightforward to show that $\preceq_K$ is a partial order. Note that the case where each decision maker has complete system knowledge and complete policy knowledge is the maximal element in this partial ordering. (See definition 2.4.)

3 Dynamic Programming Formulation

In this section we describe the dynamic programming formulation for centrally designed systems. For the rest of this section we will assume that all the random variables $\{Z_t\}$ are finite valued random variables. This allows the interchange of infimizations and expectations in the Bellman recursion to be well-defined. See [BS] for more general conditions that allow this interchange.

By Bellman's principle of optimality we may write the optimization in equation (1) as

$$\begin{aligned} &\inf_{P\in\mathcal{M}} E_P\left(f(Z_1,...,Z_T)\right) \\ &= \inf_{\{Q(Z_{j_1}|z^{j_1-1})\}} E_P\Bigg[\cdots \inf_{\{Q(Z_{j_{L-1}}|z^{j_{L-1}-1})\}} \\ &\qquad E_P\left[\inf_{\{Q(Z_{j_L}|z^{j_L-1})\}} E_P\left[f(Z_1,...,Z_T) \mid z^{j_L-1}\right] \mid z^{j_{L-1}-1}\right]\cdots \mid z^{j_1-1}\Bigg]. \end{aligned}$$

Let us examine the j_lth infimization. The decision maker at time j_l needs to compute

$$\inf_{\{Q(Z_{j_l}|z^{j_l-1})\}} E_P\left[\cdots \inf_{\{Q(Z_{j_L}|z^{j_L-1})\}} E_P\left[f(Z_1,...,Z_T) \mid z^{j_L-1}\right]\cdots \mid z^{j_l-1}\right].$$

Recall the information pattern for decision maker j_l is specified by the set J_{j_l}. The decision maker has knowledge only of the signals in $\{z_t : t \in J_{j_l}\}$. The j_lth infimization can be rewritten as

$$\inf_{\{Q(Z_{j_l}|z^{j_l-1})\}} E_P\left\{E_P\left[\cdots \inf_{\{Q(Z_{j_L}|z^{j_L-1})\}} E_P\left[f(Z_1,...,Z_T) \mid z^{j_L-1}\right]\cdots \mid z^{j_l-1}\right] \,\middle|\, \{z_t : t \in J_{j_l}\}\right\}.$$

In order to compute the inside expectations we need to know the conditional measure $P\left(Z^T \mid z^{j_l-1}\right)$ and in order to compute the outside expectation we need to know the conditional measure $P\left(Z^{j_l-1} \mid \{z_t : t \in J_{j_l}\}\right)$. We treat each case now:

(1) $P\left(Z^T \mid z^{j_l-1}\right) = Q(Z_{j_l} \mid z^{j_l-1}) \otimes \left(\otimes_{t=j_l+1}^{T} P\left(Z_t \mid z^{t-1}\right)\right)$. This product consists of system kernels and decision kernels. The decision maker has complete system knowledge. Furthermore since dynamic programming is a backward recursion the decision maker knows what the future polices will be also. Thus the decision maker knows the conditional probability $P\left(Z^T \mid z^{j_l-1}\right)$.

(2) In order to compute $P\left(Z^{j_l-1} \mid \{z_t : t \in J_{j_l}\}\right)$ we will first compute $P\left(Z^{j_l-1}\right)$. Now $P\left(Z^{j_l-1}\right) = \otimes_{t=1}^{j_l-1} P\left(Z_t \mid z^{t-1}\right)$. This product consists of past system kernels and past decision kernels. The decision maker has complete system knowledge. But unlike the case of future decision kernels it cannot compute the past decision kernels. To get around this problem we assume that the decision maker has access to the past decision kernels. Specifically we assume that the state of the dynamic programming recursion at time j_l consists of the pair

$$\left(\{z_t : t \in J_{j_l}\},\ \left\{Q\left(Z_j \mid z^{j-1}\right) \,:\, j = j_1, ..., j_{(l-1)}\right\}\right).$$

Now the decision maker can compute $P\left(Z^{j_l-1}\right)$. Consequently it can compute $P\left(Z^{j_l-1} \mid \{z_t : t \in J_{j_l}\}\right)$.

Note that at the end of the dynamic programming recursion decision maker j_1 will have chosen a policy kernel $Q^*(Z_{j_1} \mid z^{j_1-1})$. (Where * means optimal.) The optimal policy for decision maker j_2 depends on $Q^*(Z_{j_1} \mid z^{j_1-1})$. But this is part of j_2's information. Thus we see that at the end of the backward dynamic programming recursion we need to make a forward pass to substitute in the past optimal policies.

We have just shown that centrally designed systems with arbitrary information patterns can be solved via dynamic programming. Conceptually the dynamic program is straightforward. In practice though it may be very complicated to implement. Furthermore even if the objective function is convex the intermediate optimizations in the above dynamic program may be non-convex. See for example Witsenhausen's counterexample paper. [Wit3]

If the information pattern of decision maker j_l contains the whole past, i.e. $J_{j_l} = \{1, ..., j_l - 1\}$, then the decision maker does not need to know the past decision kernels. This follows because $P(Z^{j_l-1} \mid z^{j_l-1})$ is just a Dirac measure. This is a special case of Witsenhausen's work on policy independence of conditional expectations. [Wit2] This is also the traditional case, i.e. full state observation, dealt with in dynamic programming.

At the other extreme we can imagine a case where the information pattern for decision maker j_l is empty: $J_{j_l} = \emptyset$. This means that there is no signal feedback to the decision maker. In this case the decision maker needs to know all of the past decision kernels.

There are many cases in between these two extreme cases. Clearly there is a trade-off between signal knowledge, as captured by the information pattern, and system and policy knowledge. We have already shown that there exist natural partial orders on the information pattern, system knowledge, and policy knowledge.

We now continue example 2.1 of the partially observed Markov decision problem.

Example 2.1 continued *In this case the information pattern for decision maker U_t is (Y^t, U^{t-1}). Thus the infimization in the Bellman recursion requires us to compute*

$$P(X^t, Y^t, U^{t-1} \mid y^t, u^{t-1}).$$

As is well known, if the system dynamics are Markov and the cost additive then the probability of the current state given the past information is a sufficient statistic for the problem. Thus we need only compute $P(X_t \mid y^t, u^{t-1})$. But this is the usual filter for estimating the state. In this case we do not need to augment the state of the dynamic programming recursion with information about the past policies.

3.1 Difficulties with Decentralized Design

We have shown that dynamic programming can be used to design systems in a centralized way. Designing systems decentrally is much harder. This is due in part because there is no obvious analog of dynamic programming to apply here. In the centrally designed case the decision makers are coordinated in their design effort. There is an inherent game-theoretic issue when the decision makers don't have complete system or policy knowledge. It is our belief, though, that the above formulation may help in the characterization and solution of certain classes of decentrally designed distributed systems. But we leave this for future work.

There is the possibility that the different decision makers can learn the other decision makers' policies over time. This can occur, for example, if each decision maker knows the objective function and has the proper information pattern. The proper information pattern allows the decision maker the opportunity to learn another decision maker's policy through input/output data. The knowledge of the objective function and an assumption of rational design can also help the decision maker learn the other decision makers' policies. We leave the possibility of learning to future work.

There are three different kinds of knowledge in distributed systems. They are the information pattern, the system and policy knowledge, and the knowledge of the objective. Said succinctly they are knowledge of the signals, the system, and the goal. Each type of knowledge or lack of knowledge can contribute a different kind of complexity to the design.

Many distributed systems are designed in a centralized manner. We have shown that the issue of nontraditional information pattern has been, at least conceptually, taken care of provided one is allowed to centrally design the system.

Thus one may choose to redefine our popular notion of a distributed system to not be one with nontraditional information pattern but to be one that is designed decentrally. That is a system where the decision makers do not have complete system or policy knowledge.

The third sort of complexity occurs if one assumes that each decision maker has different knowledge about the control objective or have different control objectives altogether. This added complexity will not be dealt with here.

In summary there are three kinds of knowledge: signal knowledge in the form of an information pattern, system and policy knowledge, and control objective knowledge.

4 Summary

In this paper we introduced a general formulation for modeling distributed control problems. We examined the roles that information patterns and system and policy knowledge play in this framework. We showed that centrally designed systems with arbitrary information pattern can be solved, in principle, via dynamic programming.

We view a model as the set of all measures that complete a partially specified joint measure. And we view control as a selection of one of these consistent joint measures. Another way to view control is as the interconnection, under differing information patterns, of different stochastic kernels representing the system.

References

[BS] D. Bertsekas and S. Shreve, *Stochastic Optimal Control: the Discrete Time Case.* New York, Academic Press: 1978.

[Tat] S. Tatikonda, *Control under Communication Constraints.* M.I.T. Ph.D. thesis, September 2000.

[Wit] H. Witsenhausen, "Separation of Estimation and Control for Discrete Time Systems." Proceedings of the IEEE, Volume 59, No. 11, November 1971.

[Wit2] H. Witsenhausen, "On Policy Independence of Conditional Expectations." Information and Control, Volume 28, pp. 65-75, 1975.

[Wit3] H. Witsenhausen, "A Counterexample in Stochastic Optimum Control." SIAM J. Control, volume 6, pp. 131-147, 1968.

On Rate-Distortion Theory as a Convex Selection Problem*

Ioannis Kontoyiannis†
Division of Applied Math
Brown University
Providence, RI 02906
`yiannis@dam.brown.edu`
`www.dam.brown.edu/people/yiannis`

Abstract

We cast the problem of lossy data compression as a convex selection problem. Roughly speaking, there is a precise correspondence between compression algorithms and probability distributions (or measures) on the reproduction alphabet, and finding good codes corresponds to selecting good measures. We characterize the best achievable compression performance of lossy compression algorithms operating at a fixed distortion level. First we consider sources with completely arbitrary distributions. We give direct and converse coding theorems, emphasizing: (a) non-asymptotic results, and (b) results with probability one (as opposed to in expectation). Then we turn to the simplest class of random sources, namely stationary memoryless sources. We consider mixture distributions corresponding to mixture codebooks over the class of all i.i.d. codebooks, and we show that it is possible to universally encode any memoryless source with pointwise rate-redundancy of approximately $(d/2)\log n$ bits, where d is the dimension of the simplex of probability distributions over the reproduction alphabet.

1 Introduction

Let $\{X_n\ ;\ n \geq 1\}$ be a random source with values in the (general) source alphabet A.[1] Let P_n denote the marginal distribution on A^n of the block $X_1^n = (X_1, X_2, \ldots, X_n)$. For any given data string $x_1^n = (x_1, x_2, \ldots, x_n) \in A^n$ produced by this source, our goal is to find an "accurate" representation of x_1^n by a string $y_1^n = (y_1, y_2, \ldots, y_n)$ taking values in the reproduction alphabet $\hat{A}$. The accuracy or "distortion" between two such strings is measured by a family of single-letter distortion measures

$$\rho_n(x_1^n, y_1^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} \rho(x_i, y_i),$$

---

*Research supported in part by NSF Grant No. CCR-0073378.

†Permanent address: Department of Statistics, Purdue University, 1399 Mathematical Sciences Building, W. Lafayette, IN 47907-1399.

[1] To be precise, we assume that A is a complete separable metric space, equipped with its associated Borel σ-field $\mathcal{A}$. Accordingly, all measures we consider on A (or A^n) are Borel measures on $(A, \mathcal{A})$ (or $(A^n, \mathcal{A}^n)$, respectively).

where $\rho : A \times \hat{A} \to [0, \infty)$ is a fixed (measurable) distortion function. For the sake of simplicity, we assume throughout that ρ is bounded and that the reproduction alphabet $\hat{A}$ is finite (although for many of our results these assumptions are unnecessary).

We consider *variable-length codes operating at a fixed distortion level,* that is, codes C_n defined by triplets (B_n, ϕ_n, ψ_n) where:

(*a*) B_n is a subset of $\hat{A}^n$ called the *codebook*;

(*b*) $\phi_n : A^n \to B_n$ is the *encoder* or *quantizer*;

(*c*) $\psi_n : B_n \to \{0,1\}^*$ is a prefix-free representation of the elements of B_n by finite-length binary strings.

For a distortion level $D \geq 0$, we say the code $C_n = (B_n, \phi_n, \psi_n)$ *operates at distortion level* D if it encodes each source string with distortion D or less:

$$\rho_n(x_1^n, \phi_n(x_1^n)) \leq D, \qquad \text{for all } x_1^n \in A^n.$$

The quantity of interest is the description length of a code C_n, expressed in terms of the length function $\ell_n : A^n \to \mathbb{N}$:

$$\ell_n(x_1^n) = \text{ length of } [\psi_n(\phi_n(x_1^n))] \quad \text{(in bits)}.$$

The main theoretical issue is to characterize what is the best achievable compression performance of such codes. Shannon [10] gave a general characterization in terms of the rate-distortion function. Here we take a slightly different point of view and rephrase this problem as a convex selection problem. This approach has its roots in the earlier work of Kieffer [6].

In the next section we will show that there is a precise correspondence between compression algorithms and probability distributions (or measures) on the reproduction alphabet, so that constructing a good code can be reduced, at least conceptually, to picking a good probability measure. This correspondence leads to a characterization of the achievable performance of compression algorithms not in terms of the rate-distortion function, but in terms of a related quantity $K_n(D)$.

For arbitrary sources, in Section 2 we give direct and converse coding theorems, emphasizing non-asymptotic results, and results with probability one (as opposed to in expectation).

In Section 3 we restrict our attention to independent and identically distributed (i.i.d.) sources, and we examine the performance of mixture distributions, corresponding to mixture codebooks, over the class of all i.i.d. codebooks. In this setup, we show that it is possible to universally encode any memoryless source with pointwise rate-redundancy of approximately $(d/2) \log n$ bits, where $d = |\hat{A}| - 1$.

2 Arbitrary Sources

Let $B(x_1^n, D)$ denote the distortion-ball of radius D centered at $x_1^n \in A^n$:

$$B(x_1^n, D) \triangleq \{y_1^n \in \hat{A}^n : \rho(x_1^n, y_1^n) \leq D\}.$$

By a *sub-probability measure* Q_n on $\hat{A}^n$ we mean a positive measure with total mass $0 < Q_n(\hat{A}^n) \leq 1$.

For an arbitrary source $\{X_n\}$ as above, the following result demonstrates the correspondence between sequences of codes C_n operating at distortion level D, and sequences of measures Q_n on the reproduction spaces $\hat{A}^n$.

Theorem 1 (Codes-Measures Correspondence) [8]

i. *For any code* $C_n = (B_n, \phi_n, \psi_n, \ell_n)$ *operating at distortion level* D*, there is a sub-probability measure* Q_n *on* $\hat{A}^n$ *such that*

$$\ell_n(X_1^n) \geq -\log Q_n(B(X_1^n, D)) \quad \textit{bits, w.p.1}$$

ii. *For any "admissible" sequence sub-probability measures* $\{Q_n\}$ *(see below) there is a sequence of codes* $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ *operating at distortion level* D *such that*

$$\ell_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + 5\log n \quad \textit{bits, eventually, w.p.1}$$

[Above and throughout the paper, log denotes the logarithm taken to base 2.] A sequence $\{Q_n\}$ of such sub-probability measures is called *admissible* if and only if there is a finite constant R such that

$$\limsup_{n\to\infty} -\frac{1}{n}\log Q_n(B(X_1^n, D)) \leq R < \infty \quad \text{bits, w.p.1}$$

The requirement that the measures $\{Q_n\}$ be admissible is essentially equivalent to the natural requirement that the random codebooks generated according to these measures do not have an infinite rate.

Proof outline: For i., given a code $C_n = (B_n, \phi_n, \psi_n, \ell_n)$, let

$$Q_n(y_1^n) \triangleq \begin{cases} 2^{-\ell_n(y_1^n)} & \text{if } y_1^n \in B_n \\ 0 & \text{otherwise.} \end{cases}$$

The fact that Q_n is a sub-probability measure follows from Kraft's inequality. For any $x_1^n \in A^n$,

$$\begin{aligned} \ell_n(x_1^n) &= \ell_n(\phi_n(x_1^n)) \\ &= -\log Q_n(\phi_n(x_1^n)) \\ &\geq -\log Q_n(B(x_1^n, D)) \quad \text{bits.} \end{aligned}$$

The second inequality follows from the fact that C_n operates at distortion level D. Part ii. follows by a random coding argument: For each n, construct a random codebook according to the distribution Q_n. See [8] for the details. □

In view of the above result, in order to understand the achievable performance of codes we can concentrate on codes with codelengths of the form:

$$\ell_n(X_1^n) \approx -\log Q_n(B(X_1^n, D)) \quad \text{bits.}$$

In particular, our problem reduces to:

1. Picking the "best" $Q_n = Q_n^*$, and
2. Understanding the behavior of

$$\ell_n^*(X_1^n) = -\log Q_n^*(B(X_1^n, D)).$$

It is worth comparing the above setup with the corresponding problem in the lossless case. There, *all* codes have length functions that are *exactly* of the form $\ell_n(X_1^n) = -\log Q_n(X_1^n)$ (to within 1 bit), and the optimal Q_n is always the source distribution P_n.

Next, for each $D \geq 0$ we define

$$K_n(D) = K_n(P, D) = \inf_{Q_n} E_{P_n}\{-\log Q_n(B(X_1^n, D))\}$$

where the infimum is over all sub-probability measures Q_n, and we let Q_n^* denote a probability measure achieving the above infimum. Alternatively, $K_n(D)$ can be defined by

$$K_n(D) = \inf_{(X_1^n, Y_1^n)} I(X_1^n; Y_1^n) \tag{1}$$

where the infimum is over all jointly distributed random vectors (X_1^n, Y_1^n) such that $X_1^n \sim P_n$ and $\rho_n(X_1, Y_1^n) \leq D$ with probability one.

The next theorem gives finite-n bounds on the performance of arbitrary codes $C_n = (B_n, \phi_n, \psi_n, \ell_n)$.

Theorem 2 (Non-Asymptotic Bounds) [8]

Assume $K_n(D) < \infty$. *Then:*

i. *For any code* $C_n = (B_n, \phi_n, \psi_n, \ell_n)$ *operating at distortion level* D

$$E[\ell_n(X_1^n)] \geq K_n(D) \geq R_n(D) \quad \textit{bits}$$

ii. *For any* $Q_n \neq Q_n^*$:

$$\Pr\{-\log Q_n(B(X_1^n, D)) \leq -\log Q_n^*(B(X_1^n, D)) - K\} \leq 2^{-K}$$

Proof outline: Part i. is obvious in view of Theorem 1 and the relation (1). Part ii. follows from the definition of $K_n(D)$ and the generalized Kuhn-Tucker conditions given in [2]. □

The result of part ii. above is somewhat striking. It states that, for any fixed n, there is an optimal code operating at distortion level D (the code corresponding to Q_n^*) with the following property: The probability that any other code beats the optimal one by K or more bits, is at most 2^{-K}! For a detailed example see [7, Section IV].

Next, as a "sanity check," we show that for stationary ergodic sources our results asymptotically coincide with the classical ones.

Theorem 3 (Ergodic Sources)

Suppose that the source $\{X_n\}$ *is stationary and ergodic. Then the following two limits exist*

$$K(D) = \lim_{n\to\infty} \frac{1}{n} K_n(D) \qquad R(D) = \lim_{n\to\infty} \frac{1}{n} R_n(D),$$

and moreover

$$\lim_{n\to\infty} \; -\frac{1}{n} \log Q_n^*(B(X_1^n, D)) = K(D) = R(D) \;\; \textit{bits/symbol, w.p.1}$$

Proof outline: The existence of $R(D)$ is well-known [10][3]. The existence of $K(D)$ follows by subadditivity [8]. From their definitions, it is clear that $K(D) \geq R(D)$. Conversely, since there are codes that achieve $R(D)$ with probability one, from the asymptotic results of Theorem 4 below it follows that we must also have $K(D) \leq R(D)$. Finally, Kieffer [6] has shown that $-(1/n) \log Q_n^*(B(X_1^n, D)) \to K(D)$ with probability one. □

As a simple application of the codes/measures identification, we consider the problem of combining a finite number of codes into one by taking a mixture. Suppose that we have a collection of measures (or codebooks, or codes)

$$Q^{(0)} = Q^*,\ Q^{(1)},\ Q^{(2)}, \ldots, Q^{(k-1)}$$

on $\hat{A}^n$, where Q^* is in some sense the "best" code, or our "target" code. Using the (uniform) finite mixture

$$Q_M(y_1^n) \triangleq \frac{1}{k} \sum_{i=0}^{k-1} Q^{(j)}(y_1^n)$$

as a codebook distribution, the following proposition says that we can always do as well as the best code for an additional price of no more than $\log k$ bits. The proof is trivial.

Proposition (Finite Mixtures)
For any n:

$$-\log Q_M(B(X_1^n, D)) \leq -\log Q^*(B(X_1^n, D)) + \log k \ \textit{ bits, w.p.1.}$$

Next we give asymptotic upper and lower bounds on the compression performance of an arbitrary compression algorithm on an arbitrary source, in terms of the measures $\{Q_n^*\}$.

Theorem 4 (Asymptotic Bounds)
i. *For any sequence of codes* $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ *operating at distortion level* D,

$$\ell_n(X_1^n) \geq -\log Q_n^*(B(X_1^n, D)) - 2\log n \quad \textit{bits, eventually, w.p.1}$$

ii. *There is a sequence of codes* $\{C_n^* = (B_n^*, \phi_n^*, \psi_n^*, \ell_n^*)\}$ *operating at distortion level* D, *such that*

$$\ell_n^*(X_1^n) \leq -\log Q_n^*(B(X_1^n, D)) + 4\log n \quad \textit{bits, eventually, w.p.1}$$

3 Memoryless Sources

Next we consider the simplest class of sources, stationary memoryless sources. Assume that the random variables $\{X_n\}$ are i.i.d., each X_i having distribution P on A. For a fixed distortion level D let (X^*, Y^*) achieve the rate-distortion function

$$R(D) \ = \ I(X^*; Y^*),$$

and write Q^* for the distribution of the achieving Y^*.

Our next result says that, in the case of i.i.d. sources, it suffices to consider codebooks generated by product distributions on $\hat{A}^n$. Theorem 5 is an analog of Theorem 4 with the measures Q_n^* replaced by the product measures $(Q^*)^n$. Its proof is rather technical, and can be found in [8].

Theorem 5 (I.I.D Codes for I.I.D. Sources) [8]
i. *For any sequence of codes* $\{C_n = (B_n, \phi_n, \psi_n, \ell_n)\}$ *operating at distortion level* D,

$$\ell_n(X_1^n) \geq -\log(Q^*)^n(B(X_1^n, D)) - 2\log n \quad \textit{bits, eventually, w.p.1}$$

ii. *There is a sequence of codes* $\{C_m^* = (B_n^*, \phi_n^*, \psi_n^*, \ell_n^*)\}$ *operating at distortion level* D, *such that*

$$\ell_n^*(X_1^n) \leq -\log(Q^*)^n(B(X_1^n, D)) + 4\log n \quad \textit{bits, eventually, w.p.1}$$

In view of Theorem 5, we can restrict our attention to product distributions on $\hat{A}^n$. Following the corresponding developments in the lossless compression literature, we consider codes based on mixtures of these distributions (see, e.g., Davisson's classic paper [5] or the more recent work of Clarke and Barron [4]).

Let π be a prior distribution on the $(|\hat{A}| - 1)$-dimensional simplex of probability distributions Q on $\hat{A}$, and assume for simplicity that π has a density $\rho = \rho(Q)$ with respect to Lebesgue measure on the simplex.

For each n, we define the mixture distribution $Q_n^{(M)}$ on $\hat{A}^n$ by:

$$Q_n^{(M)}(y_1^n) = \int_{\text{all } Q} Q^n(y_1^n)\rho(Q)\, dQ \quad \text{for } y_1^n \in \hat{A}^n.$$

Theorem 6 (Universality of I.I.D. Mixtures) [9]
If the prior density $\rho(Q) > 0$ *for all* Q, *then as* $n \to \infty$:

$$-\frac{1}{n}\log Q_n^{(M)}(B(X_1^n, D)) \;=\; -\frac{1}{n}\log(Q^*)^n(B(X_1^n, D)) + o(1) \quad \textit{bits/symbol, w.p.1}$$

Theorem 6 says that mixture codes are universal over the class of memoryless sources. Its proof follows along the same lines as the corresponding lossless result in Barron's thesis [1]. Our next result shows that the pointwise redundancy achieved by such codes is (asymptotically) no more than

$$\frac{(|\hat{A}| - 1)}{2}\log n \text{ bits.}$$

Theorem 7 (Pointwise Redundancy of I.I.D. Mixtures) [9]
Assuming that the optimal reproduction distribution Q^* *is unique, that the density* $\rho(Q)$ *is positive for all* Q, *and under some additional smoothness conditions, as* $n \to \infty$ *we have:*

$$-\log Q_n^{(M)}(B(X_1^n, D)) \;\leq\; -\log(Q^*)^n(B(X_1^n, D)) + \frac{|A| - 1}{2}\log n + O(1) \quad \textit{bits, w.p.1}$$

The proof of Theorem 7 is based on a standard argument using Laplace's method of integration, and closely parallels the proof of the corresponding result in Clarke & Barron [4] in the lossless case.

Remarks

1. *Pointwise Redundancy = Average Redundancy.* Recall [11] that the best universal redundancy rate in expectation is:

$$E[\ell_n(X_1^n)] \approx nR(D) + \frac{|A|-1}{2}\log n.$$

Therefore, Theorem 7 shows that the best pointwise redundancy and the best average redundancy are asymptotically equal (to first order).

2. *Second-Order Pointwise Universality.* Recall [8] that even when the source is known the pointwise redundancy of the *best* code is of order $O(\sqrt{n})$:

$$-\log(Q^*)^n(B(X_1^n, D)) \approx nR(D) + \sum_{i=1}^{n} f(X_i) + O(\log n).$$

Therefore, Theorem 7 shows that it is possible to achieve not only the best rate, but also the best second order redundancy *universally.*

References

[1] A.R. Barron. *Logically Smooth Density Estimation.* PhD thesis, Dept. of Electrical Engineering, Stanford University, 1985.

[2] R. Bell and T.M. Cover. Game-theoretic optimal portfolios. *Management Sci.*, 34(6):724–733, 1988.

[3] T. Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall Inc., Englewood Cliffs, NJ, 1971.

[4] B.S. Clarke and A.R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inform. Theory*, 36(3):453–471, 1990.

[5] L.D. Davisson. Universal noiseless coding. *IEEE Trans. Inform. Theory*, 19(6):783–795, 1973.

[6] J.C. Kieffer. Sample converses in source coding theory. *IEEE Trans. Inform. Theory*, 37(2):263–268, 1991.

[7] I. Kontoyiannis. Model selection via rate-distortion theory. In *Proceedings of the 34th Annual Conference on Information Sciences and Systems, CISS 2000*, Princeton, NJ, March 2000.

[8] I. Kontoyiannis. Pointwise redundancy in lossy data compression and universal lossy data compression. *IEEE Trans. Inform. Theory*, 46(1):136–152, January 2000.

[9] I. Kontoyiannis and J. Zhang. Arbitrary source models and Bayesian codebooks in rate-distortion theory. *In preparation*, 2000.

[10] C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec.*, part 4:142–163, 1959. Reprinted in D. Slepian (ed.), *Key Papers in the Development of Information Theory*, IEEE Press, 1974.

[11] E.-h. Yang and Z. Zhang. The redundancy of source coding with a fidelity criterion – Part III: Coding at a fixed distortion level with unknown statistics. *Preprint*, 1999.

New Results in Universal Coding of Simple Non-Stationary Source Classes*

Gil I. Shamir[†] and Daniel J. Costello, Jr.[‡]
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556, U.S.A.
Emails: gshamir@nd.edu[†], Daniel.J.Costello.2@nd.edu[‡]

Abstract

Lower bounds on the redundancy for universal lossless coding of parametric piecewise stationary sources with slow linear changes in the statistics between stationary segments are derived for two settings. In the first, all changes are of equal known in advance duration d, and in the second they are of unknown durations.

The universal lossless coding problem of *Piecewise Stationary Sources* (PSS's) with abruptly changing statistics, where the number of changes, the statistical parameters in each *stationary segment*, and the change times are all unknown, has recently received much attention (see [2] and more). In [2], it was shown that for every code with length function $L(\cdot)$, the average universal coding *redundancy* over all sequences of n letters, drawn from almost every PSS ψ with q independent stationary segments, each governed by a k-dimensional parameter vector, between which $c \triangleq q-1$ abrupt transitions occur, is lower bounded by

$$R_n(L,\psi) \geq (1-\varepsilon)\left(\frac{1}{2}kq + q - 1\right)\frac{\log m}{n}, \tag{1}$$

where m is the average length of a segment, and $\varepsilon > 0$ can be arbitrarily small.

In this paper, we investigate how the behavior of the universal coding redundancy changes if we code sources that are still piecewise stationary, but contain slow changes in statistics between the stationary segments instead of abrupt changes. We refer to such a source as a *PSS with slowly Varying Statistics* (or PSVS). Like the PSS model, the PSVS model is very useful in describing data types in real applications, but, in this case, where slow changes occur between stationary regions. We restrict the discussion to the simplest case, in which the changes in statistics are linear during the *transition intervals*. We shall investigate two settings of the problem. In one, the duration d of all statistical changes is known in advance, and in the second, no a-priori information exists on the durations of all changes, which may not be identical to all changes.

A PSVS is uniquely defined by the parameter $\psi \triangleq (\boldsymbol{\theta}, \mathbf{t}^-, \mathbf{t}^+)$. The vector $\boldsymbol{\theta} \triangleq (\theta_1, \ldots, \theta_q)$ is the set of k-dimensional parameters that govern the statistics in each of

---

*This work was supported by NSF Grant CCR00-75514 and NASA Grant NAG5-8355.

the q stationary segments. The vectors $\mathbf{t}^- \triangleq \left(t_1^-, \ldots, t_{q-1}^-\right)$ and $\mathbf{t}^+ \triangleq \left(t_1^+, \ldots, t_{q-1}^+\right)$ consist of the first and last time instants of the transition intervals between stationary segments, respectively. The source parameters change linearly between θ_i and θ_{i+1} from time t_i^- to t_i^+. The vector $\mathbf{t} \triangleq (t_1, \ldots, t_{q-1})$ contains the mean points of the transition intervals.

The *redundancy* of a code with length function $L(\cdot)$ for n-sequences governed by ψ is

$$R_n(L, \psi) \triangleq \frac{1}{n} E_\psi L(X^n) - H_\psi(X^n), \tag{2}$$

where X^n is a random sequence, E_ψ is the expectation for the given PSVS, and H_ψ is the per-letter average entropy of ψ. By the *redundancy-capacity theorem* (see [1]), if we partition the *significant portion* Λ_ε of a class Λ into disjoint countable sets, each of at least M *distinguishable* sources, then

$$R_n(L, \psi) \geq (1 - \varepsilon) \frac{\log M}{n} \tag{3}$$

for every code $L(\cdot)$ and almost every source $\psi \in \Lambda$. The significant portion Λ_ε of Λ is a subset of Λ whose relative volume in Λ, under the assumption that the sources are uniformly distributed in Λ, goes to 1 as $n \to \infty$. A set of sources is distinguishable if the probability that a sequence x^n generated by a source ψ in the set appears to be generated by another source $\psi' \neq \psi$ also in the set goes to zero as $n \to \infty$.

We first assume that $q = 2$ and $\boldsymbol{\theta}$ is known. In the first setting, only t_1 is unknown. Assuming $d \to \infty$, we can let t_1 take on any point in a *grid* with spacing $d^{0.5(1+\varepsilon)}$, for $\varepsilon > 0$ arbitrarily small, and still distinguish between different sources ψ. This results in redundancy lower bounded by $(1 - \varepsilon)(\log n - 0.5 \log d)/n$ for most sources with known $\boldsymbol{\theta}$ and a single transition. Taking $d = m^\alpha$ for some constant α, $1 > \alpha > 0$, and extending this result for any q and for unknown $\boldsymbol{\theta}$ using the same techniques as in [2] results in

$$R_n(L, \psi) \geq (1 - \varepsilon) \left[\frac{kq}{2} + (q-1)\left(1 - \frac{\alpha}{2}\right)\right] \frac{\log m}{n} \tag{4}$$

for every code $L(\cdot)$ and almost every PSVS ψ with q changes of duration $d = m^\alpha$. The bound of (1) becomes a special case of (4), when $\alpha \to 0$.

Assuming $\boldsymbol{\theta}$ and t_1 are known and only d is unknown for $q = 2$ leads to a redundancy of at least $0.5(1 - \varepsilon)(\log d)/n$ for a single unknown transition duration, using the same techniques as above. Combining the last two results for a single transition results in a cost of essentially $\log n$ extra code bits if both the time and duration of a single transition are unknown. Generalizing this result, we obtain a lower bound similar to that of (1) for PSVS's. This result is surprising, since it shows that the redundancy of the general class of PSVS's that contains transitions of all possible durations is asymptotically lower bounded similarly to that of its subclass of PSS's with only abrupt transitions.

References

[1] N. Merhav and M. Feder, "A strong version of the redundancy-capacity theorem of universal coding," *IEEE Trans. Inform. Theory*, Vol. 41, pp. 714-722, May 1995.

[2] G. I. Shamir and D. J. Costello, Jr., "On the redundancy of universal lossless coding for general piecewise stationary sources," in *Proceedings of 2000 IEEE International Symposium on Information Theory*, Sorrento, Italy, p. 326, June, 2000.

On Universal Compression of Multi-Dimensional Data Arrays Using Self-Similar Curves*

Tsachy Weissman   Shie Mannor

Abstract

The LZ-based scheme for the lossless compression of multi-dimensional data arrays using self-similar curves is revisited. The asymptotic universality of this compression scheme is established for the stochastic setting in which the data array is assumed to be a realization of a stationary Markov Random Field (MRF). The existence of universal codes for the lossy compression of MRFs is also established. In particular, it is argued that the rate-distortion function of any stationary and ergodic MRF is achieved by combining a self-similar scan with Ziv's universal encoder for lossy compression.

1 Introduction

The rapidly evolving information age, invoked by the colossal developments in telecommunications and in the satellite, cellular, and networking technologies, ensues the pressing need for efficiently representing images and video signals. Practical modern compression techniques (cf. [20] [22] and references therein) are observed to be efficient and relatively robust for the storage and transmission of images and video signals.

Contrary to the abundance of reported work in the signal-, image-, and video- processing communities (cf. [13] [19] and references therein), the problem of efficiently compressing images and video signals has been given relatively little attention from an information-theoretic perspective. One of the main reasons for this is that, while in many applications involving the compression of time-series the data can be assumed to be generated by a probabilistic source belonging to a well-defined family (e.g. Markov sources or unifilar finite-state sources), in most practical situations involving images and video signals the issues related to the modeling of the data as generated by probabilistic sources are much more controversial.

Nevertheless, triggered by the seminal work of the Geman brothers [9], which established the relevance of the Bayesian setting and, in particular, of MRFs to the analysis and the processing of images, the last decade has witnessed an abundance of research work involving the modeling of images using MRFs. This approach has been proven quite fruitful in many applications related to filtering, image enhancement, and contour detection (cf. [19] and references therein). In addition, although seldom does an image or a video signal arising in applications seem like a realization of an MRF, on a local

---

*The experimental results related to this work were obtained in the Vision Research and Image Sciences Laboratory at the Technion.

Both authors are with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel. {tsachy, shie}@tx.technion.ac.il.

"microscopic" scale it is a well-known empirical fact that most types of images and video signals exhibit a neighboring behavior which can naturally be statistically modeled by an MRF with a simple (usually of bounded range) neighboring structure.

In light of all this, we believe that the stochastic setting, specifically that where the data is assumed to be a realization of an MRF, is relevant and constructive in the context of image and video compression. This belief is the motivation for this work, in which we establish the existence of universal encoders for the lossless and lossy compression of multi-dimensional data arrays which are realizations of stationary MRFs. Universality is achieved with respect to (w.r.t.) the class of all stationary MRFs whose associated potential is summable (a subset of this class, of particular interest, is the MRFs with a finite neighboring structure often used in image-processing applications: 4-nearest neighbors, 8-nearest neighbors, etc.).

The paper is organized as follows. In Section 2, a precise description of the probabilistic model assumed is given and a few known facts from this setting are recalled. Section 3 is dedicated to the lossless case in which we establish the asymptotic universality of Lempel and Ziv's algorithm for the compression of multi-dimensional data arrays (shown in [14] to be universal in an "individual picture" setting) generated by MRF sources in an exponential sense. Section 4 is dedicated to the lossy case, where it is argued that by employing Ziv's lossy encoder [23] on the space filling curve used for the lossless case one is guaranteed to achieve the rate-distortion function whenever the multi-dimensional data is generated by a stationary and ergodic source. Section 5 contains some concluding remarks.

2 The Model and Preliminaries

Let A be a finite alphabet and let $\mathcal{A}$ be the σ-algebra of all subsets of A. A *realization* or *configuration* over $\mathbb{Z}^d$ is an element $\omega = \{x_i\}_{i\in\mathbb{Z}^d}$ of the measurable space of configurations $(\Omega, \mathcal{F}) = (A, \mathcal{A})^{\mathbb{Z}^d}$. Let $\mathcal{M}(\Omega)$ denote the space of probability measures on Ω, and denote by $\mathcal{M}_s(\Omega)$ the subspace consisting of all elements of $\mathcal{M}(\Omega)$ that are invariant under all shifts $\theta_i : \Omega \to \Omega$, $i \in \mathbb{Z}^d$, where $(\theta_i\omega)(j) = \omega(j+i)$. Any element of $\mathcal{M}(\Omega)$ ($\mathcal{M}_s(\Omega)$) will be referred to as a *random field* (*stationary random field*) or, equivalently, as a *source* (*stationary source*).

For any positive integer n, let V_n be the cube of all $i = (i_1, \dots, i_d) \in \mathbb{Z}^d$ with all coordinates nonnegative and strictly less than n. For any $V \subset \mathbb{Z}^d$ define $D(V)$ as the interior diameter of V: $D(V) \overset{\text{def}}{=} \sup\{r : \exists c \text{ s.t. } B(c,r) \subseteq V\}$, where $B(c,r)$ denotes the closed ball of radius r centered at c under the l_1-norm of $\mathbb{Z}^d$. Let further $\mathcal{R}_\square$ denote the system of all d-dimensional rectangular boxes of the form $V = \mathbb{Z}^d \cap \prod_{k=1}^d [m_k, n_k]$ with $m_k, n_k \in \mathbb{Z}$, $m_k \le n_k$. For $V \subseteq \mathbb{Z}^d$, we denote by $\mathcal{F}_V$ the σ-field on Ω generated by the projections $\omega \to \omega(i)$, $i \in V$, and by ω_V the restriction of $\omega \in \Omega$ to V. Let $\mathcal{V}$ denote the class of finite subsets of $\mathbb{Z}^d$. For any source $Q \in \mathcal{M}(\Omega)$ we denote by Q_V the marginal distribution of Q on A^V, by $Q_V^*(\cdot|\omega)$ the regular *conditional distribution* on Ω given $\mathcal{F}_{V^c}$ at the point $\omega \in \Omega$, and by $Q_V(\cdot|\omega)$ its marginal distribution on A^V.

An *interaction potential* is defined by a family $\phi = \{\phi_V, V \in \mathcal{V}\}$ of maps $\phi_V : A^V \to \mathbb{R}$. It is called *stationary* if $\phi_V(\omega_V) = \phi_{V+i}((\theta_i^{-1}\omega)_{V+i})$ for all $V \in \mathcal{V}$, $\omega \in \Omega$, and $i \in \mathbb{Z}^d$. We henceforth assume a stationary interaction potential which, in addition, satisfies the summability condition $\sum_{V\in\mathcal{V}:0\in V} \|\phi_V\| < \infty$, where the norm is the supremum norm. A source $Q \in \mathcal{M}(\Omega)$ is a *Gibbs measure* with respect to ϕ if $Q_V(\cdot|\omega)$ can be chosen such that for all $\omega, \omega' \in \Omega$ and $V \in \mathcal{V}$

$$Q_V(\omega_V|\omega') = (Z_V(\omega'))^{-1} \exp\left(-U_V(\omega_V|\omega')\right), \tag{1}$$

where

$$U_V(\omega_V|\omega') \stackrel{\text{def}}{=} \sum_{B: B\cap V \neq \phi} \phi_B(\eta_B), \tag{2}$$

with $\eta \in \Omega$ defined by $\eta(i) \stackrel{\text{def}}{=} \omega(i)$ for $i \in V$ and $\eta(i) \stackrel{\text{def}}{=} \omega'(i)$ for $i \in V^c$, and where $Z_V(\omega')$ is the normalizing factor, namely

$$Z_V(\omega') \stackrel{\text{def}}{=} \sum_{\eta \in A^V} \exp\left(-U_V(\eta|\omega')\right). \tag{3}$$

We let $\mathcal{G}(\phi)$ denote the class of all Gibbs measures with respect to the potential ϕ (note that, under our assumptions of a finite alphabet and a summable potential, $\mathcal{G}(\phi)$ is guaranteed to not be empty cf., e.g., [4]). Although some pathological examples exist of elements of $\mathcal{M}_s(\Omega)$ for which a ϕ satisfying (1) does not exist, the Hammersley-Clifford theorem [11, Sec. 2.2] assures us that essentially any MRF is a Gibbs measure w.r.t. some potential. Therefore, there is no loss in assuming that our data source is a Gibbs measure.

For any $Q \in \mathcal{M}(\Omega)$ and $V \in \mathcal{V}$, we let

$$H_V(Q) \stackrel{\text{def}}{=} -\sum_{\eta \in A^V} Q_V(\eta) \log Q_V(\eta) \tag{4}$$

denote the entropy of Q_V. We define the *entropy rate* of the stationary source $Q \in \mathcal{M}_s(\Omega)$ by

$$H(Q) \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{|V_n|} H_{V_n}(Q), \tag{5}$$

where $|\cdot|$ denotes cardinality and the limit is guaranteed to always exist [10]. The terminology "entropy rate" is justified by the fact that $H(Q)$ always satisfies

$$H(Q) = \inf_{V \in \mathcal{V}} \frac{1}{|V|} H_V(Q) \tag{6}$$

which, in turn, implies that $H(Q)$ is the fundamental limitation on the compression ratio achievable for the source Q. An additional characterization of $H(Q)$, which further justifies this terminology, is the conditional entropy of a single point in the lattice, conditioned on all points in its "lexicographic past". To be more precise, we recall the lexicographic order of $\mathbb{Z}^d$ by writing $i < j$ whenever there exists some $1 \leq k \leq d$ such that $i_k < j_k$ and $i_l = j_l$ for $1 \leq l < k$. Letting $V^*(0) = \{i \in \mathbb{Z}^d : i < 0\}$ (the "lexicographic past" of the origin) one can show that

$$H(Q) = -E_Q\left(\sum_{a \in A} Q_{\{0\}}(a|\mathcal{F}_{V^*(0)}) \log Q_{\{0\}}(a|\mathcal{F}_{V^*(0)})\right), \tag{7}$$

where E_Q denotes expectation with respect to the measure Q. Proofs of these and additional interesting facts concerning $H(Q)$ can be found in [10, Ch. 15]. It is an unfortunate fact that even for the simplest MRFs (e.g. a 4-nearest neighbor field with $A = \{0, 1\}$ and $d = 2$) a closed expression for $H(Q)$ is not known.

For $V \in \mathcal{V}$ and $\mu, \nu \in \mathcal{M}_s(\Omega)$ let

$$D_V(\mu\|\nu) \stackrel{\text{def}}{=} \sum_{\eta \in A^V} \mu_V(\eta) \log \frac{\mu_V(\eta)}{\nu_V(\eta)} \tag{8}$$

denote the *relative entropy* between μ_V and ν_V. Finally, we let

$$D(\mu\|\nu) \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{|V_n|} D_{V_n}(\mu\|\nu), \tag{9}$$

where the limit exists (cf. [5, 4.25]).

Finally, we recall that the empirical field of a Gibbs measure satisfies a Large Deviations Principle (LDP) [6, 18]. Specifically, let $Q \in \mathcal{M}_s(\Omega)$ be a stationary Gibbs measure corresponding to some stationary and summable interaction potential. Define the empirical field ρ_{V_n} as the random element of $\mathcal{M}(\Omega)$ given by

$$\rho_{V_n}(\omega) = \frac{1}{n^d} \sum_{i \in V_n} \delta_{\theta_i \omega}\,, \tag{10}$$

with δ_ω the unit point mass at ω. The sequence of distributions $Q \circ \rho_{V_n}^{-1}$ of ρ_{V_n} under Q satisfies an LDP, where the rate function is given (essentially, cf. [6] for the precise statement) by the relative entropy $D(\cdot\|Q)$. Using the contraction principle [3, Section 4.2], one can obtain an LDP for the empirical distributions

$$R_{V_n}(\omega) = \frac{1}{n^d} \sum_{i \in V_n} \delta_{\omega(i)} \tag{11}$$

with the rate function

$$I_0(p) = \inf_{\nu \in \mathcal{M}_s(\Omega): \pi(\nu) = p} D(\nu\|Q), \tag{12}$$

where $\pi : \mathcal{M}(\Omega) \to \mathcal{M}(A)$ associates with each $\nu \in \mathcal{M}(\Omega)$ its marginal distribution $\nu_0 = \pi(\nu)$ at the origin $0 \in \mathbb{Z}^d$.

3 Lossless Compression

In this section we show that the scheme of Lempel and Ziv [14] for the compression of multi-dimensional data arrays is asymptotically universally optimal under the stochastic setting of the present framework. For concreteness and simplicity of notation we treat the two-dimensional case $d = 2$, as did [14]. All the results carry over straightforwardly to data arrays of any finite dimension.

To comply with the notation of [14], for any $\omega \in \Omega$ and $N \in \mathbb{N}$ we let $I_N = I_N(\omega) = \omega_{V_N}$ and we agree to "flip" the labeling of the "x-axis" around the origin so that I_N is the $N \times N$ picture occupying the upper-*left* corner of the lower-*right* quadrant for the realization ω. Let SS denote the self-similar Peano-Hilbert curve [15], [8] (starting at $(-1, -1)$ and filling the lower-right quadrant). For any $k \geq 1$ let $SS(I_{2^k})$ denote the sequence in $A^{2^{2k}}$ of the 2^{2k} symbols read from I_{2^k} under SS. Let further $L_{LZ}(I_{2^k})$ be the length function of Lempel and Ziv's [14] code for the compression of the picture I_{2^k}, namely, $L_{LZ}(I_{2^k})$ is the number of bits associated with the Lempel-Ziv codeword [24] of the sequence $SS(I_{2^k})$.

Now, as was shown in [14, Eq. 12], it easily follows from the structure of SS that for all $\omega \in \Omega$, $k \geq 1$, and $1 \leq j \leq k$ we have

$$\hat{H}_m(SS(I_{2^k})) \leq H_q^*(I_{2^k}) + \frac{2}{q^2 \log|A|}, \quad m = q^2 = 2^{2j}, \tag{13}$$

where $H_q^*(I_{2^k})$ is the normalized empirical entropy of the picture I_{2^k} as induced by the empirical distribution of its $q \times q$ subpictures and where $\hat{H}_m(SS(I_{2^k}))$ is the normalized mth-order empirical entropy of the *sequence* $SS(I_{2^k})$ as induced by its subsequences of length m. Now, combining Huffman's coding scheme for input blocks of length m (clearly implementable by a finite-state encoder) with [24, Th. 2], we obtain

$$\frac{1}{2^{2k}} L_{LZ}(I_{2^k}) \leq \hat{H}_m(SS(I_{2^k})) + \frac{1}{m} + \delta_m(k), \tag{14}$$

where $\delta_m(k)$ is deterministic (i.e., does not depend on I_{2^k}) and $\lim_{k\to\infty} \delta_m(k) = 0$. Combining (14) with (13) gives

$$\frac{1}{2^{2k}} L_{LZ}(I_{2^k}) \leq H_q^*(I_{2^k}) + \frac{C}{q^2} + \delta_m(k), \tag{15}$$

where $C = \frac{2}{\log|A|} + 1$. The idea in the remainder of this section is to use the LDP for the empirical field of a Gibbs measure and show that, for large k, the right-hand side of inequality (15) can, with high probability, be replaced by the entropy rate of the source, and to characterize the exponential rate in which the probability that this new inequality be violated decays.

Fix an arbitrary $\varepsilon > 0$ and $B \in (H(Q), \log|A|)$. Take now $q = 2^j$ sufficiently large such that $(1/q^2)H_{V_q}(Q) + C/q^2 - H(Q) < \varepsilon/2$ (this inequality clearly holds for all sufficiently large j by the definition of $H(Q)$). Now take k sufficiently large such that $\delta_m(k) < \varepsilon/2$. By inequality (15) we have

$$Q\left\{\frac{1}{2^{2k}} L_{LZ}(I_{2^k}) > B\right\} \leq Q\left\{H_q^*(I_{2^k}) > B - \varepsilon\right\}. \tag{16}$$

Taking k to infinity we obtain

$$\begin{aligned} \limsup_{k\to\infty} \frac{1}{2^{2k}} \log Q\left\{\frac{1}{2^{2k}} L_{LZ}(I_{2^k}) > B\right\} &\leq \limsup_{k\to\infty} \frac{1}{2^{2k}} \log Q\left\{H_q^*(I_{2^k}) > B - \varepsilon\right\} \qquad (17)\\ &= \limsup_{k\to\infty} \frac{1}{2^{2k}} \log Q\left\{\frac{1}{q^2} H(R_{q,2^{k-j}}(\omega)) > B - \varepsilon\right\}, \end{aligned}$$

where we define

$$R_{q,n}(\omega) = \frac{1}{n^2} \sum_{i \in V_n} \delta_{(\theta_{q \cdot i}\omega)_{V_q}}. \tag{18}$$

In words, $R_{q,n}(\omega) \in \mathcal{M}(A^{q^2})$ is the empirical measure over A^{q^2} as induced by the n^2 $q \times q$ nonoverlapping subpictures comprising $I_{q \cdot n}$. Note that $R_{q,n}$ can be viewed as the marginal empirical distribution as defined in (11) for an MRF over the super-alphabet A^{q^2}. Therefore, using the large deviations upper bound for the marginal distribution R_{V_n}, as stipulated in the previous section to hold with the rate function (12), for a field over the super-alphabet A^{q^2}, one can obtain (the details are straightforward and are omitted due to space limitations):

$$\limsup_{k\to\infty} \frac{1}{n^2} \log Q \left\{ \frac{1}{q^2} H(R_{q,n}(\omega)) > B - \varepsilon \right\} \leq - \inf_{\nu \in \mathcal{M}_s(\Omega): \frac{1}{q^2} H_{V_q}(\nu) \geq B - \varepsilon} q^2 \cdot D(\nu \| Q). \quad (19)$$

We therefore have

$$\begin{aligned} &\limsup_{k\to\infty} \frac{1}{2^{2k}} \log Q \left\{ \frac{1}{q^2} H(R_{q,2^{k-j}}(\omega)) > B - \varepsilon \right\} \\ &= \frac{1}{q^2} \limsup_{k\to\infty} \frac{1}{(2^{k-j})^2} \log Q \left\{ \frac{1}{q^2} H(R_{q,2^{k-j}}(\omega)) > B - \varepsilon \right\} \\ &\leq - \inf_{\nu \in \mathcal{M}_s(\Omega): \frac{1}{q^2} H_{V_q}(\nu) \geq B - \varepsilon} D(\nu \| Q). \end{aligned} \quad (20)$$

Combining this with (17) we obtain

$$\limsup_{k\to\infty} \frac{1}{2^{2k}} \log Q \left\{ \frac{1}{2^{2k}} L_{LZ}(I_{2^k}) > B \right\} \leq - \inf_{\nu \in \mathcal{M}_s(\Omega): \frac{1}{q^2} H_{V_q}(\nu) \geq B - \varepsilon} D(\nu \| Q). \quad (21)$$

Finally, since q and $\varepsilon > 0$ are arbitrary, we minimize the right-hand side of (21) by letting $q \to \infty$ and $\varepsilon \searrow 0$. Using the definition of the entropy rate (5), its property (6), and the continuity of $D(\cdot \| Q)$ (the details are straightforward and, again, are omitted for brevity) leads to the following result.

Theorem 1 *Let Q be any Gibbs measure associated with a stationary and summable potential. We then have for all $B \in (H(Q), \log |A|)$*

$$\limsup_{k\to\infty} \frac{1}{2^{2k}} \log Q \left\{ \frac{1}{2^{2k}} L_{LZ}(I_{2^k}) > B \right\} \leq - \inf_{\nu \in \mathcal{M}_s(\Omega): H(\nu) \geq B} D(\nu \| Q). \quad (22)$$

It will be interesting to find out whether, analogously to the results of [16] for the one-dimensional case and finite-state sources, the right-hand side of (22) is the best achievable exponent by any compression scheme, even one which is tailored for the source Q.

It can be shown that the right-hand side of (22) is strictly negative for every $B \in (H(Q), \log |A|)$. Therefore, using the Borel-Cantelli lemma (for an upper bound) and the converse to coding theorem (for a lower bound) yields

Corollary 2 *For any source Q as in Theorem 1*

$$\limsup_{k\to\infty} \frac{1}{2^{2k}} L_{LZ}(I_{2^k}) = H(Q) \quad Q - a.s. \quad (23)$$

4 Lossy Compression

We dedicate this section to briefly argue that universal *lossy* compression schemes exist for the setting of the present paper, namely that where the multi-dimensional data is the realization of an MRF. In particular, we shall argue that by scanning the data using, e.g., the Peano-Hilbert curve, and then encoding the data via Ziv's universal lossy encoder [23], one is guaranteed to achieve the fundamental compression limit, namely, the rate-distortion function of the source.

We assume that our multi-dimensional data is generated by the stationary and ergodic (cf., [10, Ch. 14] or [11, Sec. 3.1] for a precise definition, which is a natural extension of the one-dimensional case) source $P \in \mathcal{M}_s(\Omega)$. Let now $d(\cdot,\cdot) : A \times A \to [0,\infty)$ be a given one symbol distortion function. For any $V \in \mathcal{V}$, a *code* is a deterministic map $C_V : A^V \to A^V$, and

$$d_{C_V} \stackrel{\text{def}}{=} \frac{1}{|V|}\sum_{i\in V} E_P[d(X_i,(C_V(\{X_i\}_{i\in V}))_i)]$$

denotes the average distortion per symbol when the code C_V is used. Letting $|C_V|$ denote the cardinality of the range of C_V, we define the rate of the code C_V by $R_{C_V} = \frac{1}{|V|}\log|C_V|$. Using the one symbol distortion function $d(\cdot,\cdot)$, we denote the corresponding average distortion on a set $V \in \mathcal{V}$ by setting

$$d^V(\{x_i\}_{i\in V},\{y_i\}_{i\in V}) \stackrel{\text{def}}{=} \frac{1}{|V|}\sum_{i\in V} d(x_i,y_i). \tag{24}$$

We let the distortion of a probability measure Q on $A^V \times A^V$ be

$$d^V_Q = \int_{A^V\times A^V} d^V(\mathrm{x},\mathrm{y})Q(d\mathrm{x},d\mathrm{y}). \tag{25}$$

The mutual information associated with the measure Q having marginals Q_x, Q_y on A^V is defined by

$$I(Q) \stackrel{\text{def}}{=} D(Q\|Q_x\times Q_y) = \int_{A^V\times A^V}\log(\frac{dQ}{dQ_x\times Q_y})dQ. \tag{26}$$

We define the rate distortion function of the source P on the set V by

$$R_V(D) \stackrel{\text{def}}{=} \inf_{\{Q\in\mathcal{M}(A^V\times A^V):d^V_Q\le D,Q_x=P_V\}}\frac{1}{|V|}I(Q). \tag{27}$$

Finally, we define the rate distortion function for the source P by the prescription

$$R(D) \stackrel{\text{def}}{=} \lim_{n\to\infty} R_{V_n}(D), \tag{28}$$

whenever the limit exists. The definition (28) is justified by Propostion 3 below, the proof of which we now outline. The main idea is the sub-additivity of $R_V(D)$ in V, namely, the fact that for any $V_1, V_2 \in \mathcal{V}$ such that $V_1 \cap V_2 = \emptyset$ we have

$$R_{V_1\cup V_2}(D) \le \frac{|V_1|}{|V_1\cup V_2|}R_{V_1}(D) + \frac{|V_2|}{|V_1\cup V_2|}R_{V_2}(D). \tag{29}$$

To see this, note that by letting $Q^{V_1}_{y|x}$ and $Q^{V_2}_{y|x}$ denote the channels achieving the infima in $R_{V_1}(D)$ and $R_{V_2}(D)$ respectively (verification of the fact that these infima are achieved in our finite-alphabet context is straightforward) and further letting $Q^{V_1}_{y|x} \times Q^{V_2}_{y|x}$ denote the product channel obtained from $Q^{V_1}_{y|x}$ and $Q^{V_2}_{y|x}$ then

$$\begin{aligned} I\left((Q^{V_1}_{y|x}\times Q^{V_2}_{y|x})\times P_{V_1\cup V_2}\right) &\le I(Q^{V_1}_{y|x}\times P_{V_1}) + I(Q^{V_2}_{y|x}\times P_{V_2}) \\ &= |V_1|R_{V_1}(D) + |V_2|R_{V_2}(D), \end{aligned} \tag{30}$$

where the inequality is established using the observation that, under $(Q^{V_1}_{y|x} \times Q^{V_2}_{y|x}) \times P_{V_1 \cup V_2}$, $H((Y_{V_1}, Y_{V_2})|(X_{V_1}, X_{V_2})) = H(Y_{V_1}|X_{V_1}) + H(Y_{V_2}|X_{V_2})$. Combining this with the fact that

$$d^{V_1 \cup V_2}_{(Q^{V_1}_{y|x} \times Q^{V_2}_{y|x}) \times P_{V_1 \cup V_2}} \leq D \tag{31}$$

establishes equation (29). Combining equation (29) with a multi-dimensional sub-additivity lemma (the proof of which is a generalization of that of the univariate case) and with the stationarity of P it is straightforward to arrive at the following.

Proposition 3 *For any stationary source $P \in \mathcal{M}_s(\Omega)$*

(a) The limit in (28) exists.

(b) For any sequence $\{B_n\}$ of elements of $\mathcal{R}_\Box$ satisfying $D(B_n) \to \infty$ we have $R(D) = \lim_{n\to\infty} R_{B_n}(D)$.

(c) $R(D) = \inf_{\Delta \in \mathcal{R}_\Box} R_\Delta(D)$.

Shannon's source coding theorem can readily be extended to the present setting (the proof of the one-dimensional case (cf. [2] and [7]) extends to this setting in a straightforward manner; the details are omitted due to space shortage). Namely, we have

Theorem 4 (Source Coding Theorem) *Let $P \in \mathcal{M}_s(\Omega)$ be ergodic.*

1. *Direct part: For any $D \geq 0$ and any $\delta > 0$, there exists a sequence of codes $\{C_{V_n}\}_{n=1}^{\infty}$ (which depend on the source P) with distortion at most D and rates satisfying $R_{C_{V_n}} \leq R(D) + \delta$ for all n large enough.*
2. *Converse part: For any sequence of codes $\{C_{V_n}\}_{n=1}^{\infty}$ of distortion $\leq D$ and all $\delta > 0$, $\liminf_{n\to\infty} R_{C_{V_n}} \geq R(D+\delta)$.*

Unfortunately, the analytic evaluation of $R(D)$ is quite difficult and explicit expressions are known only for the simplest MRF's (the Ising model), under Hamming distortion and for D sufficiently small [1, 12].

Consider now the following scheme for the lossy compression of any image source P at rate R: scan the data using the Peano-Hilbert curve (or its extension to more than two dimensions), and then encode the resulting one-dimensional data $SS(I_{2^k})$ by employing Ziv's universal lossy encoder [23] with rate R. Decoding will naturally be done by retrieving the lossy version of $SS(I_{2^k})$ and ordering the components back into their respective coordinates in V_{2^k}. Let $\hat{I}^R_{2^k}$ denote the reconstructed version of I_{2^k} when using this scheme. Using the fact that Ziv's lossy encoder achieves the finite-state distortion-rate function for *any* individual sequence and, hence, the distortion-rate function of any (not necessarily stationary) probabilistic sequence, one can establish the universality of the above described scheme for the lossy compression of images in the following sense (we omit the details):

Theorem 5 *Let P be an arbitrary stationary and ergodic image source. For any given rate R, using the above described lossy compression scheme (which is independent of P) we have*

$$\limsup_{k\to\infty} E_P\left[d^{(2^k)}(I_{2^k}, \hat{I}^R_{2^k})\right] = D(R), \tag{32}$$

where $D(\cdot)$ is the (possibly pseudo-) inverse function of $R(\cdot)$.

5 Discussion

The lossless and lossy compression schemes presented in sections 3 and 4 operate by using a self-similar curve to scan the multi-dimensional data and turn it into a one-dimensional sequence, on which a universal compression scheme for a "temporal" process is employed.

It should be emphasized that, in general, employing a universal coding scheme on the one-dimensional sequence resulting from the scan is not guaranteed to yield a universal coding scheme of the type considered in this work. This is because even when the multi-dimensional data is generated by a very well-behaved stationary and ergodic MRF, the sequence resulting from the scan will, in general, not even be stationary. Thus, using simplistic, lossless or lossy, universal codes (of the type described in, e.g., [17]), which heavily rely on the stationarity and ergodicity properties of the sequence, will not achieve the goal.

This point is even more consequential in the search for practical, suboptimal lossy coding schemes. While the lossless scheme of Section 3 is (putting aside issues related to the implementation of the scan) practical, the lossy one of Section 4 is not. This is because the latter is based on Ziv's lossy universal encoder which requires to perform, for each block, the prohibitively complex task of finding a code, in an exponentially large class of codes, which performs best for the block. Thus one may be willing, as is often the case in one-dimensional lossy source coding, to compromise optimality for the sake of practicality and employ suboptimal universal coding schemes on the sequence resulting from the scan. The problem with this approach is that the suboptimal universal coding schemes reported in the literature, e.g. those based on string matching [21], are tailored for a stationary and ergodic sequence and hence, as discussed above, might do very poorly. The authors are currently working on obtaining practical, suboptimal coding schemes for the present setting and on assessing their performance.

6 Acknowledgement

The authors are grateful to Neri Merhav for helpful discussions and for invaluable comments on an earlier draft of the manuscript. Helpful discussions with Ofer Zeitouni are gratefully acknowledged.

References

[1] L. A. Bassalygo and R. L. Dobrushin, "ε-Entropy of the Random Field," *Prob. Peredach. Inform.* Vol. 23, No. 1, 3-15, 1987.

[2] T. Berger, *Rate Distortion Theory.* Prentice Hall, Englewood Cliffs, NJ, 1971.

[3] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications.* 2nd ed. Springer-Verlag, New York, 1998.

[4] R. L. Dobrushin, "The Description of a Random Field by Means of Conditional Probabilities and Conditions for its Regularity," *Th. Proba. Appl.*, 13, pp. 197-224.

[5] H. Föllmer, "On Entropy and Information Gain in Random Fields," *Z. Wahrsch. verw. Gebiete*, vol. 26, pp. 207-217, 1973.

[6] H. Föllmer and S. Orey, "Large Deviations for the Empirical Field of a Gibbs Measure," *Ann. of Prob.*, vol. 16, No. 3, pp. 961-977, July 1988.

[7] R. G. Gallager, *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[8] M. Gardner, "Mathematical Games," *Sci. Amer.*, pp. 124-133, Dec. 1976.

[9] D. Geman and S. Geman, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE - PAMI*, vol. 6, pp. 721-741, 1984.

[10] H. O. Georgii, *Gibbs Measures and Phase Transitions.* Walter de Gruyter, Berlin - New York, 1988.

[11] X. Guyon, *Random Fields on a Network.* Springer-Verlag, New York, 1995.

[12] B. Hajek and T. Berger, "A Decomposition Theorem for Binary Markov Random Fields," *Ann. of Prob.*, vol. 15, No. 3, pp. 1112-1125, July 1987.

[13] Hua, Harry, and Li, *Video Data Compression for Multimedia Computing.* Kluwer academic publishers, January 1997.

[14] A. Lempel and J. Ziv, "Compression of Two-Dimensional Data," *IEEE Trans. Inform. Theory*, vol. IT-32, no. 1, pp. 2-8, Jan. 1986.

[15] B. B. Mandelbrot, *Fractals: Form, Chance, and Dimension.* Freeman, San Francisco, 1977.

[16] N. Merhav, "Universal Coding with Minimum Probability of Codeword Length Overflow," *IEEE Trans. Inform. Theory*, vol. 37, pp. 556-563, May 1991.

[17] D. L. Neuhoff and P. C. Shields, "Simplistic Universal Coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, pp. 778-781, March 1998.

[18] S. Olla, "Large Deviations for Gibbs Random Fields," *Prob. Th. Rel. Fields*, vol. 77, pp. 343-357, 1988.

[19] Seul, O'Gorman, and Sammon, *Practical Algorithms for Image Analysis.* Cambridge university press, 1999.

[20] T. Sikora, "MPEG Digital Video-Coding Standards," *IEEE Sig. Proc. Magazine*, pp. 82-100, Sept. 1997.

[21] Y. Steinberg and M. Gutman, "An Algorithm for Source Coding Subject to a Fidelity Criterion, Based on String Matching," *IEEE Trans. Inform. Theory*, vol. 39, no. 3, pp. 877-886, May 1993.

[22] M. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I Lossless Image Compression Algorithm: Principles and Standardization into JPEG-LS," Computer Systems Laboratory, HPL-98-197, Nov. 1998.

[23] J. Ziv, "Distortion-Rate Theory for Individual Sequences," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 137-143, March 1980.

[24] J. Ziv and A. Lempel, "Compression of Individual Sequences Via Variable Rate Coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530-536, Sept. 1978.

Structurally Equivalent Grammars in Lossless Data Compression*

John Kieffer
Dept. of Electrical & Computer Engineering
University of Minnesota
Minneapolis, MN 55455
kieffer@ece.umn.edu

En-hui Yang
Dept. of Electrical & Computer Engineering
University of Waterloo
Waterloo, Ontario, CA N2L 3G1
ehyang@bbcr.uwaterloo.ca

Abstract

We introduce a class of grammars such that each grammar G in the class is a context-free grammar which generates a unique derivation tree, which implies that the language $L(G)$ generated by G consists of a single string. Two such grammars are *structurally equivalent* if the derivation trees they generate, stripped of all labels, are identical. Given structurally equivalent grammars G and G^*, we define conditional entropy $H(G|G^*)$ for G given G^* in a natural way. A coding theorem is presented which states that if G^* is known to both encoder and decoder, then the production rules of any grammar G structurally equivalent to G^* can be losslessly compressed using approximately $H(G|G^*)$ code bits. This result is applied to the coding of a data string x given a data string y of the same length—one finds structurally equivalent grammars G_x, G_y such that $L(G_x) = \{x\}$, $L(G_y) = \{y\}$, and then encodes x given y using roughly $H(G_x|G_y)$ code bits. With this approach, grammar-based lossless codes can be constructed which universally code any finite-alphabet information source given any second finite-alphabet information source stationarily and ergodically correlated with the first source.

1 Introduction

In [1], a theory of grammar-based data compression was set forth in which a class of grammars called admissible grammars was used. We continue this work in the present summary, but need a stronger concept of grammar that we shall call an admissible$^+$ grammar. We define a grammar G with set of variables $V(G)$ and set of terminal symbols $T(G)$ to be *admissible*$^+$ if (1) G is a context-free grammar generating a unique derivation tree in which all elements of $V(G)$ and $T(G)$ appear as labels, and (2) for any two

---

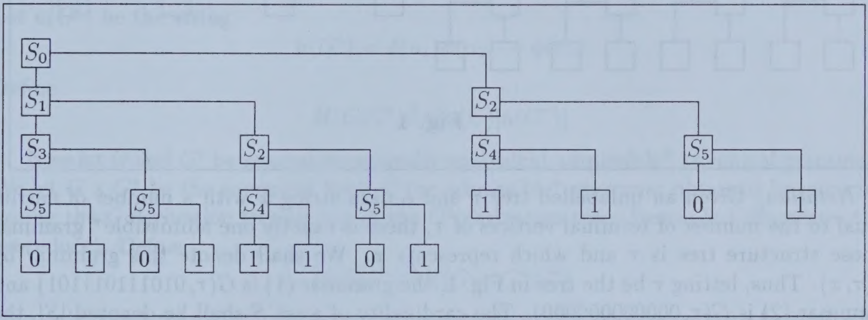
*This research was supported in part by NSF Grant CCR-9902081 and Natural Sciences and Engineering Research Council of Canada Grant RGPIN203035-98.

production rules $A \to \alpha$, $B \to \beta$ of G, $A = B$ if and only if $\alpha = \beta$. The language $L(G)$ generated by an admissible$^+$ grammar G consists of exactly one string, whose entries may be read off as the labels on the terminal vertices of the unique derivation tree, traversed in depth-first order.

For example, the grammar with production rules

$$\begin{array}{lll} S_0 \to S_1S_2 & S_1 \to S_3S_2 & S_2 \to S_4S_5 \\ S_3 \to S_5S_5 & S_4 \to 11 & S_5 \to 01 \end{array} \tag{1}$$

is admissible$^+$, and the unique derivation tree that it generates is given by



From the figure, it is clear that the data string represented by the grammar (1) is 010111011101.

A grammar-based source code [1] compresses a data string indirectly by losslessly compressing an admissible$^+$ grammar which represents the string. Whenever an admissible$^+$ grammar G is encoded, we require that G be *canonical* [1], by which we mean that the variables in $V(G)$ are ordered according to their first appearances as labels in the breadth-first traversal of the vertices of the derivation tree of the grammar (this is the *canonical ordering* of $V(G)$). In [1], we explained how to compute an entropy $H(G)$ for an admissible$^+$ canonical grammar G, such that the grammar G may be encoded into a binary codeword of length approximately $H(G)$ from which G may be losslessly recovered.

In the current paper, for certain pairs of canonical admissible$^+$ grammars (G, G^*), we shall define the *conditional entropy* $H(G|G^*)$ of the grammar G given the grammar G^*. The quantity $H(G|G^*)$ is interpreted in Theorem 1, which states that G may be encoded with the help of G^* into a binary codeword $w(G|G^*)$ of length approximately $H(G|G^*)$, and that G can be decoded from $w(G|G^*)$ with the help of G^*. (Notice that in this approach, the grammar G^* is regarded as "side-information" provided to both encoder and decoder.)

We define the *structure tree* of an admissible$^+$ grammar to be the unlabelled tree which results when the derivation tree of the grammar is stripped of all labels. In order for the conditional entropy $H(G|G^*)$ to be defined, we must require that the grammars G and G^* be *structurally equivalent*, by which we mean that G, G^* have the same structure tree. For example, the grammar

$$\begin{aligned} S_0 &\to S_1S_2 \\ S_1 &\to S_2S_2 \\ S_2 &\to S_3S_3 \\ S_3 &\to 00 \end{aligned} \tag{2}$$

and the grammar (1) are structurally equivalent because they have the same structure tree, namely, the tree in Fig. 1.

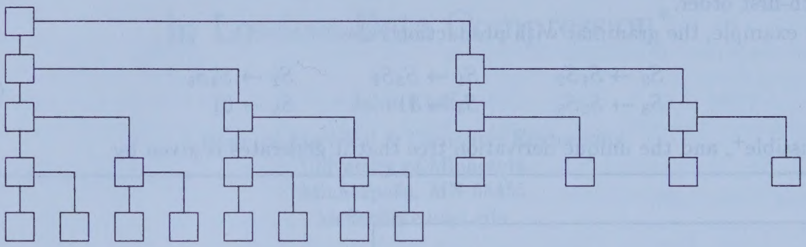


Fig. 1

Notation. Given an unlabelled tree τ and a data string x with a number of entries equal to the number of terminal vertices of τ, there is exactly one admissible$^+$ grammar whose structure tree is τ and which represents x. We shall denote this grammar by $G(\tau, x)$. Thus, letting τ be the tree in Fig. 1, the grammar (1) is $G(\tau, 010111011101)$ and grammar (2) is $G(\tau, 000000000000)$. The cardinality of a set S shall be denoted $|S|$, the length of a string x shall be denoted $|x|$, and the sum of the lengths of the right members of the production rules of an admissible$^+$ grammar G shall be denoted $|G|$.

2 Conditional Entropy

If G, G^* are admissible$^+$ grammars, we say that a mapping $\phi : V(G) \cup T(G) \to V(G^*) \cup T(G^*)$ is a *homomorphism of* G *onto* G^* if both of the following conditions hold:

(a) ϕ maps $V(G)$ onto $V(G^*)$ and $T(G)$ onto $T(G^*)$.

(b) If $A \to \alpha_1\alpha_2 \cdots \alpha_j$ is a production rule of G, then $\phi(A) \to \phi(\alpha_1)\phi(\alpha_2)\cdots\phi(\alpha_j)$ is a production rule of G^*, and every G^* production rule is mapped into in this way by at least one G production rule.

If there is a homomorphism of G onto G^*, then it is unique [2] and G, G^* are structurally equivalent. We first define conditional entropy $H(G|G^*)$ of the grammar G given the grammar G^* for the case in which there is a homomorphism of G onto G^*.

Suppose $X = x_1x_2\cdots x_j$ and $Y = y_1y_2\cdots y_j$ are strings of the same length. We define the (first order) conditional entropy $h(X|Y)$ of X given Y as follows:

$$h(X|Y) \triangleq \sum_{i=1}^{j} -\log_2 p(x_i|y_i)$$

where for each pair $(x, y) \in \{(x_1, y_1), (x_2, y_2), \cdots, (x_j, y_j)\}$,

$$p(x|y) \triangleq \frac{\mathrm{card}\{1 \le i \le j : (x_i, y_i) = (x, y)\}}{\mathrm{card}\{1 \le i \le j : y_i = y\}}$$

Let ϕ be a homomorphism from the admissible$^+$ canonical grammar G onto the admissible$^+$ canonical grammar G^*. For each variable $A \in V(G)$, let $s(A)$ be the right side of the production rule of G whose left side is A. Let the canonical ordering of the variables of G be $A_0, A_1, A_2, \cdots, A_{|V(G)|-1}$. In the string

$$s(A_0)s(A_1)\cdots s(A_{|V(G)|-1})$$

cross out the first left-to-right appearance of each variable A_i, $i \neq 0$. Let $u(G)$ be the string of length $k = |G| - |V(G)| + 1$ that is left, which we write out as

$$u(G) = u_1 u_2 \cdots u_k$$

Let $u(G^*)$ be the string

$$u(G^*) = \phi(u_1)\phi(u_2)\cdots\phi(u_k)$$

Define

$$H(G|G^*) \triangleq h(u(G)|u(G^*)) \tag{3}$$

Now let G and G^* be general structurally equivalent admissible$^+$ canonical grammars. We let $G \times G^*$ be the canonical form of the admissible$^+$ grammar obtained by superimposing the G derivation tree on top of the G^* derivation tree (Example 1 illustrates this procedure). Define

$$H(G|G^*) \triangleq H(G \times G^*|G^*) \tag{4}$$

The right hand side of (4) makes sense because there is a homomorphism from $G \times G^*$ onto G^*. If there is a homomorphism from G onto G^*, then $G \times G^* = G$ and both definitions (3) (4) give the same result for $H(G|G^*)$.

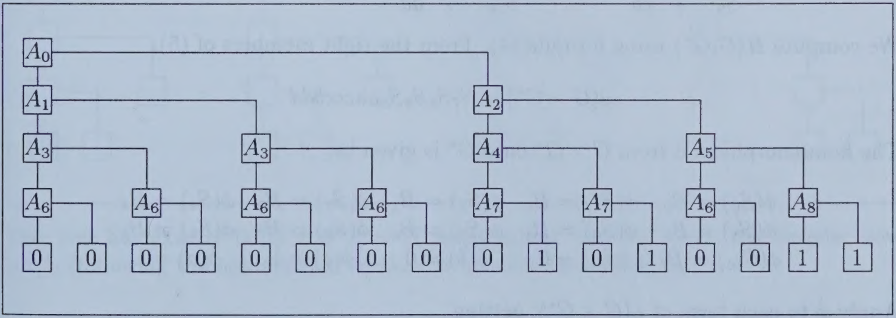
Example 1. Let G be the admissible$^+$ grammar with production rules

$$\begin{array}{lll} A_0 \to A_1A_2 & A_1 \to A_3A_3 & A_2 \to A_4A_5 \\ A_3 \to A_6A_6 & A_4 \to A_7A_7 & A_5 \to A_6A_8 \\ A_6 \to 00 & A_7 \to 01 & A_8 \to 11 \end{array}$$

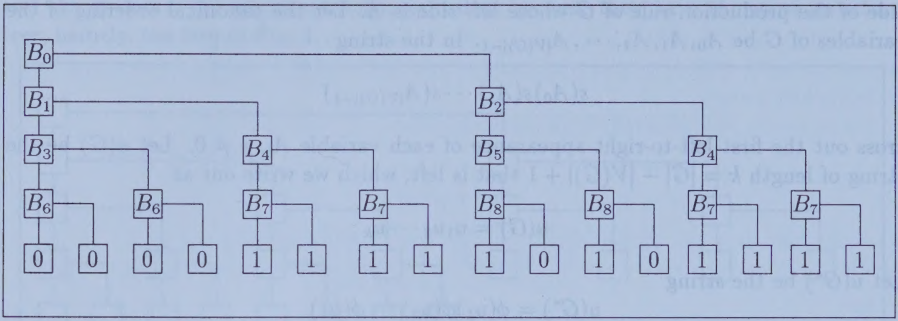
and let G^* be the admissible$^+$ grammar with production rules

$$\begin{array}{lll} B_0 \to B_1B_2 & B_1 \to B_3B_4 & B_2 \to B_5B_4 \\ B_3 \to B_6B_6 & B_4 \to B_7B_7 & B_5 \to B_8B_8 \\ B_6 \to 00 & B_7 \to 11 & B_8 \to 10 \end{array}$$

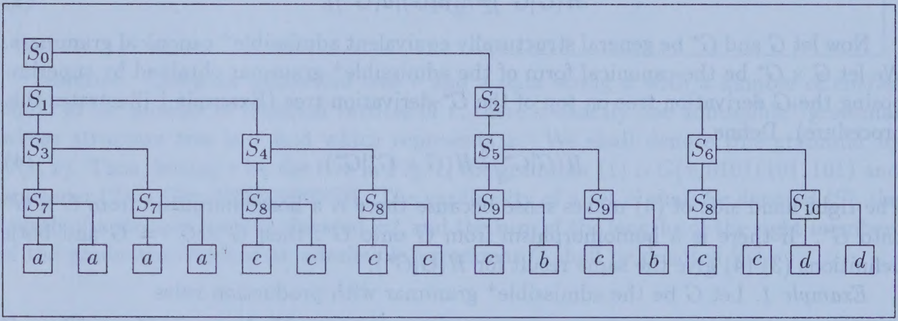
The derivation tree of G is



and the derivation tree of G^* is


B_0
B_1 B_2
B_3 B_4 B_5 B_4
B_6 B_6 B_7 B_7 B_8 B_8 B_7 B_7
0 0 0 0 1 1 1 1 1 0 1 0 1 1 1 1

Superimposing the G derivation tree on top of the G^* derivation tree, we obtain the tree


S_0
S_1 S_2
S_3 S_4 S_5 S_6
S_7 S_7 S_8 S_8 S_9 S_9 S_8 S_{10}
a a a a c c c c c b c b c c d d

where

$$\begin{array}{llll}
S_0 = (A_0, B_0) & S_4 = (A_3, B_4) & S_8 = (A_6, B_7) & a = (0,0) \\
S_1 = (A_1, B_1) & S_5 = (A_4, B_5) & S_9 = (A_7, B_8) & b = (1,0) \\
S_2 = (A_2, B_2) & S_6 = (A_5, B_4) & S_{10} = (A_8, B_7) & c = (0,1) \\
S_3 = (A_3, B_3) & S_7 = (A_6, B_6) & & d = (1,1)
\end{array}$$

The grammar which has this tree as its unique derivation tree is $G \times G^*$ and its production rules are

$$\begin{array}{lllllllll}
S_0 & \to & S_1S_2 & S_1 & \to & S_3S_4 & S_2 & \to & S_5S_6 \\
S_3 & \to & S_7S_7 & S_4 & \to & S_8S_8 & S_5 & \to & S_9S_9 \\
S_6 & \to & S_8S_{10} & S_7 & \to & aa & S_8 & \to & cc \\
S_9 & \to & cb & S_{10} & \to & dd & & &
\end{array} \tag{5}$$

We compute $H(G|G^*)$ using formula (4). From the right members of (5),

$$u(G \times G^*) = S_7S_8S_9S_8aacccbdd$$

The homomorphism ϕ from $G \times G^*$ onto G^* is given by

$$\begin{array}{lllll}
\phi(S_0) = B_0 & \phi(S_1) = B_1 & \phi(S_2) = B_2 & \phi(S_3) = B_3 & \phi(S_4) = B_4 \\
\phi(S_5) = B_5 & \phi(S_6) = B_4 & \phi(S_7) = B_6 & \phi(S_8) = B_7 & \phi(S_9) = B_8 \\
\phi(S_{10}) = B_7 & \phi(a) = 0 & \phi(b) = 0 & \phi(c) = 1 & \phi(d) = 1
\end{array}$$

Apply ϕ to each term of $u(G \times G^*)$, getting

$$u(G^*) = B_6B_7B_8B_700111011$$

This gives us:

$$
\begin{aligned}
H(G|G^*) &= h(u(G \times G^*)|u(G^*)) \\
&= h(aab|000) + h(cccdd|11111) \\
&= [-2\log_2(2/3) - \log_2(1/3)] + [-3\log_3(3/5) - 2\log_2(2/5)] = 7.61
\end{aligned}
$$

3 Coding Theorems

In our concluding section, we state our main results Theorem 1 and Theorem 2. With the concepts defined in the preceding sections, the reader can understand the statements of these results. Due to space limitations, we cannot prove these results here—we refer the reader to the paper [2] for the proofs.

We will need the following notation. We fix a finite nonempty set $\mathcal{A}$ for the rest of the summary. If G^* is an admissible$^+$ canonical grammar, we let $\mathcal{G}(\mathcal{A}|G^*)$ be the set of all admissible$^+$ canonical grammars G such that $T(G) \subset \mathcal{A}$ and such that G, G^* are structurally equivalent.

Our first result tells us that $H(G|G^*)$ is the approximate number of code bits needed to encode each $G \in \mathcal{G}(\mathcal{A}|G^*)$, provided that encoder and decoder know G^*.

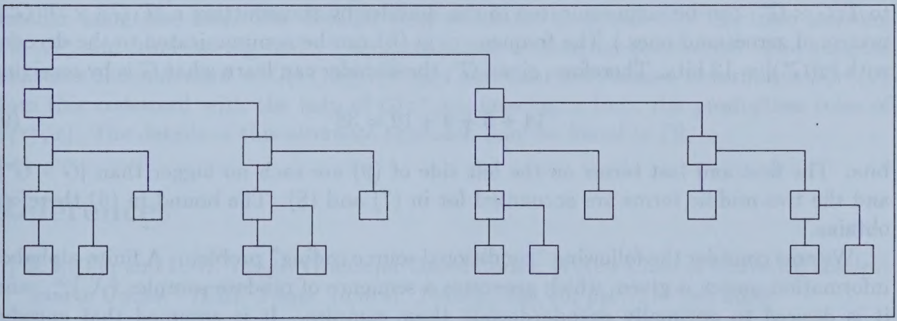
Theorem 1 *Let G^* be any admissible$^+$ canonical grammar. Then, for each grammar $G \in \mathcal{G}(\mathcal{A}|G^*)$, there is a binary string $w(G|G^*)$ of length no greater than*

$$2|G \times G^*| + |\mathcal{A}||T(G^*)| + \lceil H(G|G^*) \rceil \tag{6}$$

such that $G_1 = G_2$ whenever $G_1, G_2 \in \mathcal{G}(\mathcal{A}|G^*)$ and $w(G_1|G^*)$ is a prefix of $w(G_2|G^*)$.

Remark. In typical applications of Theorem 1, the last term $\lceil H(G|G^*) \rceil$ of (6) will dominate the other terms, so that the length of the codeword $w(G|G^*)$ will indeed be approximately equal to $H(G|G^*)$.

Example 2. We use the grammars G, G^* of Example 1 to illustrate Theorem 1. First, the encoder sends 14 bits to the encoder in order that the decoder may build the tree



This tree has exactly $|V(G \times G^*)| = 11$ nonterminal vertices, which the decoder now labels (following the nonterminal vertices in breadth-first order) as:

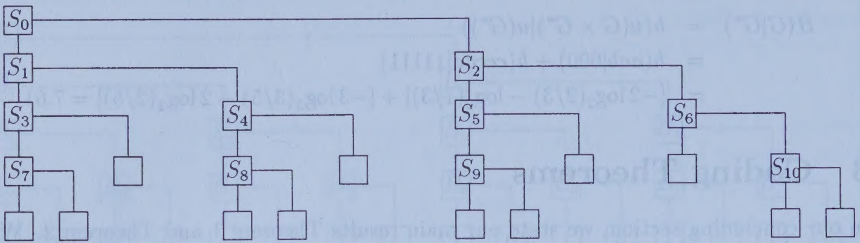


Fig. 2

The decoder will know $G \times G^*$ and then can find G once the $G \times G^*$ labels on the 12 terminal vertices in Fig. 2 (traversed in depth-first order) are determined. These labels yield a permutation of the string $u(G \times G^*)$, and the G^* labels on these same vertices yield the same permutation of the string $u(G^*)$, which the decoder knows. The encoder can communicate the required permutation of the string $u(G \times G^*)$ to the decoder via enumerative coding with

$$\lceil h(u(G \times G^*)|u(G^*)) \rceil = \lceil H(G|G^*) \rceil = 8 \tag{7}$$

bits, provided that

(a): the homomorphism $\phi : V(G \times G^*) \times T(G \times G^*) \to V(G^*) \times T(G^*)$ is known; and

(b): the frequency of each element of $V(G \times G^*) \cup T(G \times G^*)$ in $u(G \times G^*)$ is known.

The homomorphism ϕ can be communicated to the decoder with

$$|T(G)||T(G^*)| = 4 \tag{8}$$

bits. (The restriction of ϕ to $V(G \times G^*)$ is already known from the Fig. 2 tree because the decoder just has to look at the G^* labels on the nonterminal vertices; the restriction of ϕ to $T(G \times G^*)$ can be communicated to the decoder by transmitting a $|T(G)| \times |T(G^*)|$ matrix of zeroes and ones.) The frequencies in (b) can be communicated to the decoder with $|u(G^*)| = 12$ bits. Therefore, given G^*, the decoder can learn what G is by receiving

$$14 + 8 + 4 + 12 = 38 \tag{9}$$

bits. The first and last terms on the left side of (9) are each no bigger than $|G \times G^*|$, and the two middle terms are accounted for in (7) and (8). The bound in (6) therefore obtains.

We now consider the following "conditional source coding" problem. A finite-alphabet information source is given, which generates a sequence of random samples $\{X_i\}_{i=1}^{\infty}$, and it is desired to optimally encode/decode these samples. It is assumed that encoder and decoder are allowed to observe the sequence of random samples $\{Y_i\}_{i=1}^{\infty}$ generated by a second finite-alphabet information source stationarily and ergodically correlated with the first one. We can use the prefix code $G \to w(G|G^*)$ of Theorem 1 to implement a grammar-based coding scheme for this problem, in the following way. For each $n = 1, 2, \cdots$, an unlabelled tree τ_n with n terminal vertices is chosen. (The encoder and decoder agree on this choice beforehand.) With knowledge of the strings

$X^n = X_1X_2\cdots X_n$ and $Y^n = Y_1Y_2\cdots Y_n$, the encoder forms the structurally equivalent grammars $G(\tau_n, X^n), G(\tau_n, Y^n)$ and then encodes the string X^n into the binary codeword $w(G(\tau_n, X^n)|G(\tau_n, Y^n))$ with the help of the grammar $G(\tau_n, Y^n)$. With knowledge of the string Y^n, the decoder builds the grammar $G(\tau_n, Y^n)$, decodes the grammar $G(\tau_n, X^n)$ from the codeword $w(G(\tau_n, X^n)|G(\tau_n, Y^n))$ with the help of the grammar $G(\tau_n, Y^n)$, and then "grows" the string X^n using the production rules of $G(\tau_n, X^n)$. Is this grammar-based coding scheme asymptotically optimal? Let $H(X|Y)$ denote the conditional entropy rate of the process $\{X_i\}$ given the process $\{Y_i\}$, defined by

$$H(X|Y) \stackrel{\Delta}{=} \lim_{n\to\infty} n^{-1} \sum_{x\in A^n, y\in B^n} -\Pr[X^n = x, Y^n = y] \log_2 \Pr[X^n = x|Y^n = y]$$

Source coding theory tells us that our lossless coding scheme is asymptotically optimal if and only if the coding rate $n^{-1}|w(G(\tau, X^n)|G(\tau, Y^n))|$ in bits/sample converges to $H(X|Y)$. The coding theorem which follows gives us some weak requirements under which this will be true. In the parlance of source coding theory, our grammar-based coding scheme for losslessly coding one source given another is *universal*, because it achieves asymptotic optimality regardless of the source statistics.

Theorem 2 *Let A, B be finite nonempty sets. For each $n = 1, 2, \cdots$, let τ_n be a tree with n terminal vertices. Suppose that*

$$\max_{z\in(A\times B)^n} |G(\tau_n, z)| = o(n),$$

Let $\{(X_i, Y_i) : i = 1, 2, \cdots\}$ be a stationary ergodic process consisting of $A \times B$-valued random variables. Then,

$$\lim_{n\to\infty} n^{-1} H(G(\tau_n, X^n)|G(\tau_n, Y^n)) = H(X|Y) \text{ almost surely.}$$

Remark. We sketch an alternate approach in which the same tree τ_n need not be used to encode every string x in A^n given a string y in B^n. The encoder can determine a tree τ^x that depends on $x \in A^n$, provided that this tree can be communicated to the decoder using $o(n)$ bits. The decoder then builds the grammar $G(\tau^x, y)$ from knowing $y \in B^n$ and from having had τ^x communicated to him/her previously. The encoder transmits the codeword $w(G(\tau^x, x)|G(\tau^x, y))$ and then the decoder determines $G(\tau^x, x)$ from this codeword with the help of $G(\tau^x, y)$, growing x from the production rules of $G(\tau^x, x)$. The details of this alternate approach may be found in [2].

References

[1] J. Kieffer and E.-H. Yang, "Grammar-Based Codes: A New Class of Universal Lossless Source Codes," *IEEE Trans. Inform. Theory*, Vol. 46, pp. 737–754, 2000.

[2] J. Kieffer and E.-H. Yang, "Grammar Structure and Homomorphism Concepts in Grammar-Based Lossless Source Code Design," paper in preparation.

Lossless Compression of Rich Alphabet Sources via Backward processing

I. Sutskover, B. Oklander, E. Surijan and J. Ziv
Department of Electrical Engineering
Technion - I.I.T, Haifa, 32000, Israel
Email: ilan@siglab.technion.ac.il

Abstract

The problem of universal lossless compression of rich alphabet source sequences is investigated. This richness of a source is intimately related to the length of the sequence to be compressed, and thus bounds in the non-asymptotic regime are needed. In a forthcoming paper, some non-asymptotic lower bounds on the achievable compression are proven by Ziv for the class of stationary sources and a class of admissible algorithms [1]. An admissible algorithm which comes near these bounds was suggested in [2, 3]. In this paper we focus on extension of the class of admissible algorithms to allow incorporation of some prior assumptions, thus creating an "almost universal" class of algorithms. This extension is performed by introducing the method of "backward processing".

1 Introduction

Universal coding is used for the compression of source sequences when the statistics of the source are unknown. Some universal algorithms are known to have asymptotic performance that is as good as the optimal non-universal algorithm, e.g. [4, 5, 6]. However, in many cases, sequences are too "short" for such asymptotic results to apply. For example, consider a gray-level image having 512 rows and 512 columns, represented by 8 bits per pixel. The length of this image is 2^{21} bits. Assume that the image was drawn of a Markov source with a moderate order, such that every pixel depends on some nearest three pixels. Then, the number of parameters is 2^{32}, assuming Risannen's approach [6]. Obviously, the image is too short for such asymptotic results to be applicable. We would not call this image "short" was the alphabet of the source binary for example. In the later case, the number of parameters is 2^4 which is considerably smaller than the sequence's length. Hence, the shortness of a sequence is intimately related with the richness of the source alphabet. Henceforth, the term non-asymptotic results refers to results that apply to such short sequences (alternatively, such rich alphabet).

In a forthcoming paper by Ziv [1], a class of admissible sequential, universal algorithms is defined. Lower bounds for such admissible algorithms are derived, when a limited training data is used. These bounds apply to the compression of any stationary source, under one of the two assumptions to be later specified. In [3], a universal algorithm named PHZ was introduced and analyzed. It was shown that its performance comes near the lower bounds stated in [1] and refined in [2]. As common in universal coding, some cost must be "paid" for the

---

Part of this work was carried out at the Signal and Image Processing Laboratory (http://www-sipl.technion.ac.il).

unknown statistics. This paper exposes the tradeoff involved (in the non-asymptotic regime), and proposes a way to incorporate prior assumptions that do not come in a probabilistic natural way, such as "smoothness of an image".

In the next section we present the class of admissible universal algorithms, some appropriate lower bounds and the PHZ algorithm. Then, Section 3 presents the extension of this class of algorithms to algorithms that are "almost universal" using the backward processing method. Lower bounds on this class of algorithms and an upper bound on the performance of a proper version of the PHZ algorithm are given. Section 4 presents application of the new algorithm to compression of noisy images. Section 5 concludes.

2 The class of admissible algorithms and the PHZ algorithm

Let $X_{-t}^{1} = X_{-t}, \ldots, X_1$ be a sequence of random variables drawn out of a stationary source with a finite alphabet $\mathcal{A}$. We are interested in a fixed-to-variable algorithm that compresses X_1 (to be referred to as "the next symbol"). Let $Y_{-N+1}, \ldots, Y_0$ be a training data of length N. Define the random variable K_0 as the largest positive integer $k \leq t$ such that $X_{-k}^{0} = Y_{-j-k}^{-j}$ for some $0 \leq j < N-k$. We formally set $K_0 = (-1)$ if X_0 does not occur in the training data. Note that K_0 is a function of X_{-t}^{0} and Y_{-N+1}^{0} but is not a function of X_1. Define the random variable K_1 as the largest positive integer $k \leq t$ such that[1] $X_{-k}^{0}x = Y_{-j-k-1}^{-j}$ for some $0 \leq j < N-k-1$. We formally set $K_1 = (-1)$ if X_0^1 or X_1 do not occur in the training data. Again, by abuse of notation we write K_1 instead of $K_1(X_{-t}^{0}, Y_{-N+1}^{0}, X_1)$. The K_0-constraint entropy, denoted by $H(X_1|X_{-K_0-1}^{0}, Y_{-N+1}^{0})$, and the K_1-constraint entropy, denoted by $H(X_1|X_{-K_1}^{0}, Y_{-N+1}^{0}, K_1)$, are naturally defined (see [3]).

Let $Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0})$ be the probability assignment of a universal algorithm for the encoding of X_1 given the past X_{-t}^{0} and using a training data Y_{-N+1}^{0}. The class of admissible algorithms contains any universal coding algorithm that obey the following two assumptions:

Assumption 1: The algorithm is not using any phrase longer than the longest match $X_{-K_0}^{0}$ as a context for the encoding of X_1. That is, $Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0}) = Q_{Y_{-N+1}^{0}}(X_1|X_{-K_0}^{0})$.

Assumption 2: Consider the encoding probability assignment of the form

$$Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0}, K_1 = k) = \frac{Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0})}{\sum_{x:K_1(x)=k} Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0})}.$$

Then, given the value of $K_1(x)$, the algorithm is not using any phrase longer than $X_{-K_1(x)}^{0}$ as a context for the encoding of a symbol x. That is,

$$Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0}, K_1) = Q_{Y_{-N+1}^{0}}(X_1|X_{-K_1}^{0}, K_1).$$

Theorem 1 (Converse theorem [3]).

a) *For any stationary source and any admissible algorithm,*

$$\begin{aligned} -E\log_2 Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0}) &\geq H(X_1|X_{-K_1}^{0}, K_1, Y_{-N+1}^{0}) + H(K_1|X_{-K_0-1}^{0}, Y_{-N+1}^{0}) \\ &\geq H(X_1|X_{-K_1}^{0}, Y_{-N+1}^{0}) + H(K_1|X_{-K_0-1}^{0}, Y_{-N+1}^{0}) - H(K_1|X_{-K_1}^{0}, Y_{-N+1}^{0}). \end{aligned}$$

b) *If the training data is a realization independent of the sequence X_{-t}^{∞}, then*

$$\begin{aligned} &-E\log_2 Q_{Y_{-N+1}^{0}}(X_1|X_{-t}^{0}) \\ &\quad \geq H(X_1|X_{-K_1}^{0}) + H(K_1|X_{-K_0-1}^{0}, Y_{-N+1}^{0}) - H(K_1|X_{-K_1}^{0}, Y_{-N+1}^{0}). \end{aligned}$$

---

[1] The Term $X_{-k}^{0}x$ stands for the concatenation of the random phrase X_{-k}^{0} and the symbol x.

c) *If the training data is a suffix of the sequence $X^0_{-\infty}$, and the source is a Markov source with finite but unknown memory length K such that $K \leq \mathcal{O}\left(N^{\frac{\delta}{2}}\right)$, then for t < K*

$$
\begin{aligned}
&-E \log_2 Q_{Y^0_{-N+1}}(X_1|X^0_{-t}) \\
&\quad \geq H(X_1|X^0_{-K_1}) + H(K_1|X^0_{-K_0-1}, Y^0_{-N+1}) - H(K_1|X^0_{-K_1}, Y^0_{-N+1}) - \mathcal{O}\left(\frac{1}{N^\delta}\right)
\end{aligned} \quad (1)
$$

where $0 \leq \delta \leq 1$ is a parameter that depends on the source.

We observe that the memory constraint, N, need not grow exponentially with the memory length of the source, K, thus keeping us away from the classic asymptotic region.

Let y be a vector in $\mathcal{A}^\ell$ and let $w(x) \in \mathcal{A}$ denote any symbol that is different from the symbol $x \in \mathcal{A}$. The conditional recurrence time $n(x|y, Y^0_{-\infty})$ is the number of times that phrases $w(x)y$ appear in $Y^0_{-\infty}$, until the first time that a phrase xy appear. The semi-infinite sequence $Y^0_{-\infty}$ defines the direction of searching.

The PHZ algorithm [2, 3] belongs to the class of admissible algorithms defined above. It follows the following steps, for any time instant: First, it evaluates $K_1(x)$ for every $x \in \mathcal{A}$. Denote by S_k the set $\{x : K_1(x) = k\}$. All the symbols are sorted by their K_1 value so that a symbol is higher on the list when its value of K_1 is bigger. Also, symbols inside a set, S_k, are sorted with respect to their conditional recurrence time, such that a symbol is higher on the list if its recurrence time is shorter. The PHZ algorithm encodes the symbol X_1 using two indices: One that points to the correct set $S_{K_1(X_1)}$ and the other that points to the symbol X_1 within this set. The first pointer is coded using equal probability on the set of all possible values of K_1. Let $\mathcal{K} = \{k : \exists x, K_1(x) = k\}$. Then, a probability of $\frac{1}{|\mathcal{K}|}$ is used to encode the first pointer. Note that $|\mathcal{K}| \leq |\mathcal{A}|$. Let $j \in \{1, \ldots, |S_{K_1(X_1)}|\}$ denote the value of the second pointer. Then, the probability distribution used is $\frac{1/j}{\sum 1/i}$ where the sum is over all possible values of j. We refer to the latter code as *the Wyner code* [7]. This code is useful when the probabilities of the source are ordered but unknown (cf. Ryabko[8].)

Theorem 2. *Let $L_{PHZ}(X_1|X^0_{-N+1})$ denote the length function of the PHZ algorithm using a sliding window of length N. Then, for any stationary source,*

$$
E\left[L_{PHZ}(X_1|X^0_{-N+1})\right] \leq H(X_1|X^0_{-K_1}, K_1) + E\left[\log|\mathcal{K}|\right] + \rho_W \quad (2)
$$

where ρ_W stands for the redundancy of the Wyner code, and it is less than $\log_2(1 + \ln(|\mathcal{A}|))$.

Note that the lower bound (1) requires the source to be Markov, whereas the bound in Theorem 2 does not. Simulation results show that for many types of files, the set $\mathcal{K}$ is very small, and equation (2) does not differ by much from (1).

An efficient algorithm exists for the PHZ. A version of it is described in [9].

3 Incorporation of prior assumptions into universal algorithms

In many practical cases, some prior assumptions are available. When these assumptions match the model of the source, a reduction in compression rate is possible, as compared with universal coding. For example, it is common to assume that an image captured by a camera has many "smooth" parts. In [3] the class of admissible algorithms was extended to include algorithms that use a-priori assumptions in one of two ways:

Backward processing: Here, some processing of the past X^0_{-t} and the training data Y^0_{-N+1} is allowed. The new class of algorithms may use a processed context for the encoding of the symbol X_1 (which is not allowed to be processed). Such a processing function is, for example, a quantizer. Another example of an admissible processing function is a causal finite-memory filter.

Backward distortion: Here, a match between the context and a phrase in the training data is subject to a fidelity criterion. Any context that is similar to a phrase in the training data may be used, where similarity is defined by a distortion measure.

Note that both methods generate no loss. For these two generic methods, lower bounds and upper bounds may be found [2]. This paper focuses on the backward processing method, for which an efficient version of the PHZ was found. Some more motivation can be found in the following examples, which relate to lossless image compression:

Example 1. *Consider the case of a natural image accompanied by additive noise. For lossless compression of this noisy image some additional cost must be paid over the cost of the clean image. Typically, this cost will be paid twice - once due to the noisy next symbol and once due to the noisy past which dictates the statistics database for the compression of the next symbol. But do we have to pay both prices? A proper causal de-noising filter can reduce that noise in the past, enabling us to suffer only from the noise in the next symbol.*

Example 2. *In many cases, the exact value of pixels in the past of an image is not directly related to the next symbol as some approximate value of it. In these cases, a simple quantization of the past may efficiently reduce the model extra cost. Indeed, such an action was previously suggested in LOCO-I [10].*

Definition 1. *A function $q : \mathcal{A}^* \to \mathcal{B}^*$ is called an admissible processing function if it is a causal, time-invariant, Borel-measurable function having a finite-memory length.*

If a sequence X^{∞}_{-N+1} is to be compressed, where X^0_{-N+1} was already encoded at the encoder end and decoded at the decoder end, then the backward processing method applies an admissible processing function to X^0_{-N+1} yielding a new process B^0_{-N+1} (over the alphabet $\mathcal{B}$), and then encodes X_1 using the processed past B^0_{-N+1}. We may assume that the compression algorithm uses a training data (could be a sliding window) Y^0_{-N+1}, and we denote by C^0_{-N+1} the process $q\left(Y^0_{-N+1}\right)$.

We redefine the random variables K_0 and K_1.

Definition 2. *The random variable K_0 is defined as the largest positive integer $k \leq t$ such that $B^0_{-k} = C^{-j}_{-j-k}$ for some $0 \leq j < N - k$. It is formally set to -1 if B_0 is not in the training data. It should be noted that K_0 is not a function of X_1.*

Definition 3. *The random variable $K_1(x)$ is defined as the largest positive integer $k \leq t$ such that $B^0_{-k}x = C^{-j-1}_{-j-k-1}Y_{-j}$ for some $0 \leq j < N - 1 - k$. We formally set K_1 to (-1) if X_1 or B_0X_1 do not occur in the past (i.e. in Y^0_{-N+1} or its processes version respectively). It should be noted that K_1 is a function of x and the sequences X^0_{-N+1} and Y^0_{-N+1}.*

Definition 4. *$\mathcal{K}_B$ is the set of all possible values of $K_1(X_1, B^0_{-N+1})$. Note that $|\mathcal{K}_B| \leq |\mathcal{A}|$, since for any symbol x there may be only one value of $K_1(x)$.*

The new class of admissible algorithms (which now cannot be truly called universal) abides by the following two assumptions:

Assumption 1
We assume that the discussion is limited to the class of empirical conditional distributions such that, for $t \geq K_0$, $Q_{Y^0_{-N+1}}(X_1|B^0_{-t})$ is independent of $B^{-K_0-1}_{-t}$. The reasoning is that the longest event that relates to B^0_{-t} is derived from $B^0_{-K_0}$.

Note that we condition the empirical distribution on Y^0_{-N+1} in syntax, but for pattern matching purposes, a deterministic on-line processing is performed on Y^0_{-N+1} to yield C^0_{-N+1}.

Assumption 2
We consider the empirical conditional distribution of the form

$$Q_{Y^0_{-N+1}}(X_1|B^0_{-t}, K_1 = k) = \frac{Q_{Y^0_{-N+1}}(X_1|B^0_{-t})}{\sum_{X_1:K_1(X_1)=k} Q_{Y^0_{-N+1}}(X_1|B^0_{-t})} \tag{3}$$

We now assume that the discussion is limited to empirical conditional distributions such that $Q_{Y^0_{-N+1}}(X_1|B^0_{-t}, K_1 = k)$ is independent of B^{-k-1}_{-t} given that $K_1 = k$.

The followings are our main results.

Theorem 3. *For an independent database Y^0_{-N+1}, under both assumptions 1 and 2, and for any $t = 0, 1, 2, 3, \ldots$*

$$\begin{aligned} -E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-t}) &\geq H(X_1|B^0_{-K_1}, K_1, Y^0_{-N+1}) + H(K_1|B^0_{-K_0-1}, Y^0_{-N+1}) \\ &\geq H(X_1|B^0_{-K_1}) - \left[H(K_1|B^0_{-K_1}, Y^0_{-N+1}) - H(K_1|B^0_{-K_0-1}, Y^0_{-N+1})\right] \end{aligned} \tag{4}$$

If the training data is a suffix of the sequence $X^0_{-\infty}$ and the source is a Markov source with finite but unknown memory length K such that $K < \mathcal{O}\left(N^{\frac{\delta}{2}}\right)$ then as in Theorem 1, we suffer a loss of another $\mathcal{O}\left(\frac{1}{N^\delta}\right)$.

Theorem 4. *Let $L(X_1|B^0_{-N+1})$ be the length function of the code word for X_1 assigned by the sliding window PHZ algorithm using backward processing. Then*

$$E\left[L(X_1|B^0_{-N+1})\right] \leq H(X_1|B^0_{-K_1}, K_1) + E[\log|\mathcal{K}_B|] + \rho_W \tag{5}$$

where ρ_W stands for the redundancy of the Wyner code, and it is less than $\log(1 + \ln(|\mathcal{A}|))$.

The proof of Theorem 3 and the outline of the proof of Theorem 4 are given at the appendix. Complete proofs may be found in [2].

Comparing (5) to (4) reveals again the power of the PHZ scheme - the lower bound is nearly achieved. However, the effect of the backward processing method needs to be emphasized. Even if the PHZ is achieving the lower bound on algorithms using the specific backward processing function, this lower bound should now be compared to the performance of other backward processing functions or even to the performance of PHZ with no backward processing at all.

An interesting lower bound is therefore

$$\min H(X_1|B^0_{-K_1}, K_1, Y^0_{-N+1}) + H(K_1|B^0_{-K_0-1}, Y^0_{-N+1}) \tag{6}$$

where the minimization is over all admissible processing function.

The best strategy for sequentially designing the admissible processing function by means of (6) remains an open question.

Note also that no matter what strategy is adopted, or what processing function is a-priori chosen, the term in (6) is higher than or equal to the source entropy because we deal with lossless compression algorithms.

Thus, using backward processing may improve the achievable compression ratio. We would now like to briefly discuss the way by which this is done. The processing function $q(\cdot)$ may cause a rate reduction due to:

- *Intra set effect*
  The PHZ algorithm orders and codes the symbols using two indices. One index points to the set $S_{K_1(X_1)}$ which contains X_1, and the other index points to X_1 itself in this set. By intra set effect we mean the way by which the backward processing effects the coding within the set, i.e., the length assigned to the second index.

  By using a backward processing, the size of the set $S_{K_1(X_1)}$ may be reduced. A reduction in $H(X_1|B^0_{-K_1}, K_1, Y^0_{-N+1})$ is therefore possible, but not assured, since the processing of the side information may lead to a reduction in the usefulness of the side information causing an over all increase in $H(X_1|B^0_{-K_1}, K_1, Y^0_{-N+1})$. The Wyner code redundancy in the case of reduction in $|S_{K_1(X_1)}|$ is reduced, which is another cause for rate reduction.

- *Inter set effect*
  The inter set effect is the reduction of the amount of sets S_k. This effect therefore reduces the cost term $E\log|\mathcal{K}_{\mathcal{B}}|$, but the cause of this effect may also cause an inverse effect on the intra set coding. For example, by enlargement of the past alphabet (i.e. $|\mathcal{B}| > |\mathcal{A}|$), many letters are likely to have short values of K_1. Then, if the set of short K_1 is large enough, the intra set coding cost may be very high.

4 Application to compression of noisy images

We would like to elaborate on Example 1 by presenting some simulation results. Table 1 presents the performance of three algorithms: the universal PHZ algorithm described in Section 2, the Lempel-Ziv algorithm as obtained by UNIX's 'gzip' and LOCO-I [10]. All three algorithms were applied to noisy versions of the famous 'peppers' image, at 8 bpp rate and at 256 × 256 size. Raster scan is used for all algorithms. The added noise is uniform in a symmetric interval $[-\alpha, \alpha]$. The context was limited[2] to $t \leq 10$.

Noise Variance	PHZ rate [bpp]	GZIP rate [bpp]	LOCO-I rate [bpp]
0	6.6111	7.019	4.433
83	7.009	7.343	4.930
667	7.549	7.510	5.441
2250	7.986	7.588	5.878
5333	8.270	7.631	6.167
10417	8.474	7.683	6.470

Table 1: Compression results for the image "peppers".

It should be emphasized that the first two algorithms are universal, whereas the third is specially adapted to images, and therefore is not universal.

---

[2] This limitation was maintained for all PHZ simulations shown in this paper.

In order to improve the performance of PHZ for the compression of natural gray-level images, we now incorporate some prior assumptions in the form of backward processing. Figure 1 shows the neighborhood, $\{a, b, c, x\}$, of the next pixel to be encoded, y. The pixel x is the symbol that was just been encoded, i.e. X_0 in the discussion above.

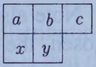


Figure 1: Pixels' locations

First we check a quantizer as a processing function. Table 2 presents the rate achieved by gathering any q consecutive gray levels, for different values of q, for the 'peppers' image with noise of variance 667. It was observed that quantization step of 7 is usually suitable for gray-level images.

q	3	5	7	9
Rate	7.242	7.199	7.202	7.209

Table 2: Performance of PHZ with different quantization levels on the 'peppers' image

Next, we add a filter. The principles of appropriate filter design were not investigated. It was observed that the filter

$$\frac{1}{9}\left(3 * Q(x) + 4 * Q(c) + Q(b) + Q(a)\right) \tag{7}$$

suits noisy images, where $Q(x)$ is the quantized value of the pixel x using a step of size 7, and the locations are depicted in Figure 1. The filter's output is the processed value of the pixel in the location x. The next symbol to encode is the pixel in location y.

We now present some comparisons. Table 3 presents the rate obtained for the 'peppers' image with different amounts of noise. These results can be compared to the first column of Table 1. The improvement gained using the backward processing method is about 0.5 bit per pixel.

Noise variance	0	83	667	2250	5333	10417
Rate	6.277	6.515	6.914	7.287	7.556	7.786

Table 3: Performance of PHZ with quantization and filtering on 'peppers' image

Table 4 compares compression on the noiseless 512 × 512 image 'Lena'.

Algorithm:	Universal PHZ	GZIP	LOCO-I	PHZ with quantization	PHZ with the filter in (7)
Rate:	6.237	6.804	4.243	6.315	6.054

Table 4: Comparison of compression methods on the 'Lena' image

Finally, Figure 2 compares the behavior of two algorithms - WinZip (a Lempel-Ziv [4] based algorithm) and PHZ with backward processing (using the processing function in (7)). Again, we simulate additive noise with different levels of variance. It is seen that the PHZ with backward processing is not as saturated at low noise variance as WinZip is. A possible explanation for this phenomenon is that PHZ with the backward processing is indeed capturing the source model of the image, so additional noise is affecting its performance. The WinZip algorithm, on the other hand, is suffering badly from noise already at low variances, so more noise is affecting it only by a little.

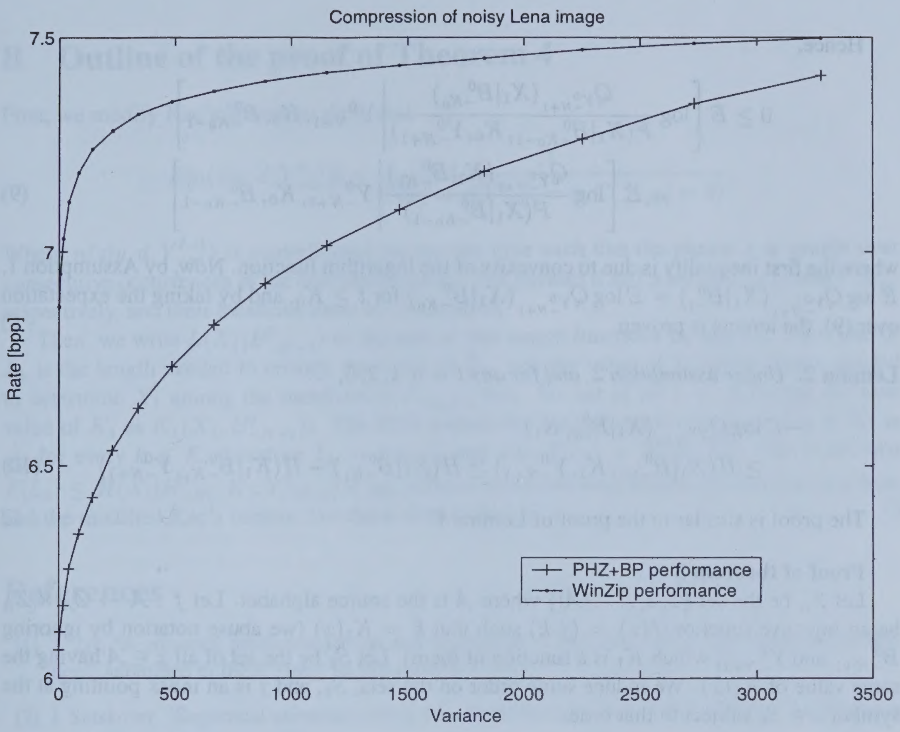


Figure 2: Compression results of the 'Lena' image using different amount of noise.

5 Conclusions

Fixed-to-Variable algorithms provide natural candidates for algorithms whose performance can be analyzed in the non-asymptotic regime. The PHZ algorithm was shown have performance close to optimal in [3, 2]. In this paper, the concept of backward processing was applied to the PHZ algorithm, in order to utilize prior assumptions about the source and match its model. It was observed that many classical signal processing tools, such as filters and quantizers, are appropriate in matching the source model for the task of lossless coding. The benefit of these tools was demonstrated on the task of lossless compression of noisy images.

A Proof of Theorem 3

Lemma 1. *Under assumption 1, and for any* $t = 0, 1, 2, 3, \ldots$

$$-E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-t}) \ge H(X_1|B^0_{-K_0-1}) \tag{8}$$

Proof: Since K_0 is a function of B^0_{-N+1} and Y^0_{-N+1} only and not a function of X_1, we get $P(X_1|B^0_{-K_0-1}, K_0, Y^0_{-N+1}) = P(X_1|B^0_{-K_0-1}, Y^0_{-N+1})$. Because X^0_{-t} is independent of Y^0_{-N+1} for any value of t (and so is B^0_{-t}) then $P(X_1|B^0_{-K_0-1}, Y^0_{-N+1}) = P(X_1|B^0_{-K_0-1})$

Hence,

$$\begin{aligned}0 &\geq E\left[\log \frac{Q_{Y^0_{-N+1}}(X_1|B^0_{-K_0})}{P(X_1|B^0_{-K_0-1},K_0,Y^0_{-N+1})}\middle| Y^0_{-N+1},K_0,B^0_{-K_0-1}\right]\\ &= E\left[\log \frac{Q_{Y^0_{-N+1}}(X_1|B^0_{-K_0})}{P(X_1|B^0_{-K_0-1})}\middle| Y^0_{-N+1},K_0,B^0_{-K_0-1}\right]\end{aligned} \tag{9}$$

where the first inequality is due to convexity of the logarithm function. Now, by Assumption 1, $E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-t}) = E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-K_0})$ for $t \geq K_0$, and by taking the expectation over (9), the lemma is proven. ☐

Lemma 2. *Under assumption 2, and for any* $t = 0, 1, 2, 3, \ldots$

$$\begin{aligned}&-E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-t},K_1)\\ &\quad\geq H(X_1|B^0_{-K_1},K_1,Y^0_{-N+1}) \geq H(X_1|B^0_{-K_1}) - H(K_1|B^0_{-K_1},Y^0_{-N+1})\end{aligned} \tag{10}$$

The proof is similar to the proof of Lemma 1. ☐

Proof of theorem 3

Let Z_A be the set $\{1,2,\ldots,|\mathcal{A}|\}$ where $\mathcal{A}$ is the source alphabet. Let $f:\mathcal{A}\to Z_A\times Z_A$ be an injective function $f(x)=(j,k)$ such that $k=K_1(x)$ (we abuse notation by ignoring B^0_{-N+1} and Y^0_{-N+1}, which K_1 is a function of them). Let S_k be the set of all $x\in\mathcal{A}$ having the same value of $K_1(x)$. We induce some order on the sets, S_k, and j is an index pointing at the symbol $x\in S_k$ subject to that order.

Then,

$$\begin{aligned}&-E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-t})\\ &\quad\geq min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-t})\\ &\quad\overset{(a)}{\geq} min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(X_1|B^0_{-K_0-1})\\ &\quad\overset{(b)}{=} min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(j(X_1),k(X_1)|B^0_{-K_0-1})\\ &\quad\geq min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(j(X_1)|k(X_1),B^0_{-K_0-1})\\ &\qquad+ min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(k(X_1)|B^0_{-K_0-1})\\ &\quad\overset{(c)}{=} min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(X_1|k(X_1),B^0_{-K_0-1})\\ &\qquad+ min_{Q_{Y^0_{-N+1}}} - E\log Q_{Y^0_{-N+1}}(k(X_1)|B^0_{-K_0-1})\\ &\quad\overset{(d)}{\geq} H(X_1|k(X_1),B^0_{-K_1},Y^0_{-N+1}) + H(k(X_1)|B^0_{-K_0-1},Y^0_{-N+1})\\ &\quad\geq H(X_1|B^0_{-K_1},Y^0_{-N+1}) + H(k(X_1)|B^0_{-K_0-1},Y^0_{-N+1}) - H(k(X_1)|B^0_{-K_1},Y^0_{-N+1})\\ &\quad\overset{(e)}{=} H(X_1|B^0_{-K_1}) + H(k(X_1)|B^0_{-K_0-1},Y^0_{-N+1}) - H(k(X_1)|B^0_{-K_1},Y^0_{-N+1})\end{aligned} \tag{11}$$

where inequality (a) is due to Lemma 1 and assumption 1, equalities (b) and (c) are because $f(x)=(j,k)$ is an injective function, inequality (d) is because of Lemma 2, assumption 2 and the fact that $k(X_1)=K_1(X_1)$, and equality (e) is due to independence of the realizations. ☐

B Outline of the proof of Theorem 4

First, we modify Kac's lemma to show that

$$E[n(z|y,d,Y_{-\infty}^{\ell-1})] \leq \frac{1}{P(Y_1^{\ell} = z|Y_{-j+1}^{0} = y, C_{-j-i+1}^{-j} = d)}.$$

Where $n(z|y,d,Y_{-\infty}^{\ell-1})$ is a conditional recurrence time such that the phrase z is sought over consecutive occurrences of the concatenation dy. The phrases d and y are in C^{0}_{-N+1} and Y^{0}_{-N+1} respectively, and their locations there are contiguous.

Then, we write $L(X_1|B^{0}_{-N+1})$ as the sum of two length functions L_k and L_x. The value of L_k is the length needed to encode the value of K_1, and the value of L_x is the length needed to determine X_1 among the members of $S_{K_1(X_1)}$ (i.e. the set of all $x \in \mathcal{A}$ having the save value of K_1 as $K_1(X_1, B^{0}_{-N+1})$). The PHZ probability assignment for the encoding of K_1 is $\frac{1}{|\mathcal{K}|}$ for every $k \in \mathcal{K}$, therefore $L_k = \log|\mathcal{K}|$ and also $E[L_k] = E[\log|\mathcal{K}|]$. The claim that $E[L_x] \leq H(X_1|B^{0}_{-K_1}, K_1, Y^{0}_{-N+1}) + \rho_W$ follows from ordering according to recurrence time and the modified Kac's lemma. For more details see [2]. □

References

[1] J. Ziv, "A universal prediction lemma and application to universal data compression and prediction", submitted to IEEE Trans. on Information Theory.

[2] I. Sutskover, "Sequential universal coding with limited training data", M.Sc. Thesis, The Technion - I.I.T, Israel, December 1999.

[3] I. Sutskover and J. Ziv, "On sequential universal coding of sequences with limited amount of training data", In proceedings of the 21st IEEE Convention of the Electrical and Electronic Engineers in Israel, pp. 365-368, Tel Aviv, Israel, 2000.

[4] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression", IEEE Trans. on Information Theory, IT-23(3):337-343, May 1977.

[5] L.D. Davisson. "Universal noiseless coding", IEEE Trans. on Information Theory, IT-19(6):783-795, November 1973.

[6] J. Rissanen. "Universal coding, information, prediction, and estimation", IEEE Trans. on Information Theory, IT-30(4):629-636, July 1984.

[7] P. Elias, "Universal codeword sets and representations of the integers", IEEE Trans. on Information Theory, IT-21(4):194-203, March 1975.

[8] B. Ya. Ryabko, "Coding of a source with unknown but ordered probabilities", Problems of information transmission, Vol. 15, No. 2, pp. 134-138, April-July 1979.

[9] Z. Cohen *et al, "On the temporal HZY compression scheme",SODA 2000, ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA USA 2000.*

[10] M.J. Weinberger et al, " LOCO-I: A low complexity, context-based, lossless image compression algorithm", In 1996 Data Compression Conference, pp. 140-149, Snowbird, Utah, USA, March 1996.

Uniformly Improving the Generalized Likelihood Ratio Test for Linear Gaussian Channels*

Elona Erez
Dept. of Electrical Engineering-Systems
Tel Aviv University
Tel Aviv, 69978, Israel
elona@eng.tau.ac.il

Meir Feder
Dept. of Electrical Engineering-Systems
Tel Aviv University
Tel Aviv, 69978, Israel
meir@eng.tau.ac.il

Abstract

We consider the decoding problem for unknown Gaussian linear channels. Important examples of linear channels are the inter-symbol interference (ISI) channel and the diversity channel with multiple transmit and receive antenna employing space time (ST) codes. Most of the solutions to this problem are based on the generalized likelihood ratio test (GLRT) decoding rule. Our work deals with a decoding algorithm that uniformly improves the exponential order of the error probability for unknown ISI channels. This exponential improvement is uniform, i.e., for all possible ISI coefficients the error probability is either smaller by an exponential factor, or, only for some, remains the same.

1 Introduction

When the communication channel is band-limited, signal transmission at a symbol rate equal or exceeding the bandwidth of the channel results in inter-symbol interference (ISI) among a number of adjacent symbols. The channel model of a Gaussian ISI channel is given by,

$$\mathcal{H}_m : y_n = \sum_{i=0}^{K-1} h_i x_{n-i}^{(m)} + z_n, \quad n = 0, 1, \ldots, N+K-1 \qquad m = 1, \ldots, M \tag{1}$$

where M is the number of codewords, $\{y_n\}_{n=0}^{N+K-1}$ are the observed data samples, $\{x_n^{(m)}\}_{n=0}^{N-1}$ are the transmitted symbols for the mth codeword, $\{h_i\}_{i=0}^{K-1}$ are the unknown ISI coefficients and z_n is white Gaussian noise with variance σ^2.

We can write (1) as :

$$H_m : \mathbf{y} = X_m \mathbf{h} + \mathbf{z} \tag{2}$$

where

$$\mathbf{y} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N+K-1} \end{bmatrix}, \mathbf{z} = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_{N+K-1} \end{bmatrix} \tag{3}$$

---

*This research is supported by the Israeli Software Radio Consortium

$$X_m = \begin{bmatrix} x_0^{(m)} & 0 & \cdots & 0 \\ x_1^{(m)} & x_0^{(m)} & \cdots & 0 \\ x_2^{(m)} & x_1^{(m)} & \ddots & 0 \\ \vdots & \vdots & \ddots & x_0^{(m)} \\ x_{N-1}^{(m)} & \vdots & \ddots & \vdots \\ 0 & x_{N-1}^{(m)} & \ddots & \vdots \\ 0 & 0 & \cdots & x_{N-1}^{(m)} \end{bmatrix}, \mathbf{h} = \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_{K-1} \end{bmatrix} \tag{4}$$

where the matrices X_m are assumed to be full rank.

Diversity channels are another class of linear channels. Diversity channels include several antennas at the transmitter and several antennas at the receiver (see [1]). The model for a diversity channel with K transmitting elements and a single receiving antenna is:

$$\begin{aligned} \mathcal{H}_m : y_n = \sum_{i=1}^{K} \alpha_i x_{n,i}^{(m)} + z_n, \quad n &= 0, 1, \ldots, N-1 \\ m &= 1, \ldots, M \end{aligned} \tag{5}$$

where $\{y_n\}_{n=0}^{N-1}$ are the observed data samples at receive antenna, $\{x_{n,i}^{(m)}\}_{n=0}^{N-1}$ are the symbols transmitted by the ith antenna for the mth codeword, α_i is the unknown fading coefficient from transmit antenna i to receive antenna and z_n is white Gaussian noise with variance σ^2. Note that for space time codes X_m in (3) can be any full rank $N \times K$ matrix. The new decoder we introduce can equally be applied to ISI channels and diversity channels with a single receive antenna.

2 GLRT Decoder for ISI Channels

We summarize the GLRT decoding rule for M codewords, which performs joint channel and data estimation. To decide which of the codewords was transmitted, for the channel in (2) we use the following criterion:

$$\hat{m} = \arg\min_m \{\min_{\mathbf{h}} ||\mathbf{y} - X_m \mathbf{h}||^2\} \tag{6}$$

Assuming that X_m are full rank matrices the least square solution for $\mathbf{h}$ is:

$$\hat{\mathbf{h}}_m = (X_m^T X_m)^{-1} X_m^T \mathbf{y} \tag{7}$$

Substituting into (6) yields the following closed form solution:

$$\begin{aligned} \hat{m} &= \arg\min_m \{||\mathbf{y} - X_m (X_m^T X_m)^{-1} X_m \mathbf{y}||^2\} & (8) \\ &= \arg\min_m \{||\mathbf{y}||^2 - \mathbf{y}^T X_m (X_m^T X_m)^{-1} X_m^T \mathbf{y}\} & (9) \end{aligned}$$

or

$$\hat{m} = \arg\max_m \{\mathbf{y}^T X_m (X_m^T X_m)^{-1} X_m^T \mathbf{y}\} \tag{10}$$

For two codewords (M=2) the separating surface of the decoder is given by:

$$S^G = \{\mathbf{v} : \mathbf{v}^T (X_1 (X_1^T X_1)^{-1} X_1^T - X_2 (X_2^T X_2)^{-1} X_2^T) \mathbf{v} = 0\}; \tag{11}$$

Define the two subspaces each of the codewords spans:

$$C_m = \{\mathbf{r}_m : \mathbf{r}_m = X_m\mathbf{h}, \mathbf{h} \in \Re^K\} \tag{12}$$

We will use these definitions in the following sections.

3 Uniformly Improving the GLRT: Motivation

We begin by considering the two codewords case. We analyze the error probability of the GLRT decoder, given an ISI coefficients vector $\mathbf{h}$. Define:

$$d_1^G(\mathbf{h}) = \min_{\mathbf{v}\in S^G}\{\|X_1\mathbf{h} - \mathbf{v}\|\} \tag{13}$$

$$d_2^G(\mathbf{h}) = \min_{\mathbf{v}\in S^G}\{\|X_2\mathbf{h} - \mathbf{v}\|\} \tag{14}$$

where $\mathbf{h}, \mathbf{X_m} (m = 1, 2)$ are defined in (2) and S^G is the separating surface of the GLRT decoder defined in (11).

The vector $\mathbf{z}$ in (2) is a white Gaussian random vector and we assume that the two messages are equiprobable. Therefore the error probability given $\mathbf{h}$ for the GLRT decoder can be approximated by:

$$P_e^G(\mathbf{h}) \approx \frac{1}{2}Q(\frac{d_1^G(\mathbf{h})}{\sigma}) + \frac{1}{2}Q(\frac{d_2^G(\mathbf{h})}{\sigma}) \tag{15}$$

Assume $d_1^G(\mathbf{h}) < d_2^G(\mathbf{h})$. The exponential order of $P_e^G(\mathbf{h})$ is given by

$$P_e^G(\mathbf{h}) \sim e^{-\frac{d_1^G(\mathbf{h})^2}{2\sigma^2}} \tag{16}$$

Define a new decoder by another separating surface S^N and respective distances

$$d_1^N(\mathbf{h}) = \min_{\mathbf{v}\in S^N}\{\|X_1\mathbf{h} - \mathbf{v}\|\} \tag{17}$$

$$d_2^N(\mathbf{h}) = \min_{\mathbf{v}\in S^N}\{\|X_2\mathbf{h} - \mathbf{v}\|\} \tag{18}$$

Suppose the new decoder maintains the following conditions:

$$\min_{m=1,2}\{d_m^N(\mathbf{h})\} \geq \min_{m=1,2}\{d_m^G(\mathbf{h})\}, \forall\mathbf{h} \in \Re^K \tag{19a}$$

$$\exists\mathbf{h} : \min_{m=1,2}\{d_m^N(\mathbf{h})\} > \min_{m=1,2}\{d_m^G(\mathbf{h})\}, \mathbf{h} \in \Re^K \tag{19b}$$

These conditions ensure that for some $\mathbf{h}$ the error of the new decoder is improved exponentially, while for the rest it remains at least the same. For the case of fading with no ISI (K=1 in (1)) such decoder was found in the Appendix of [2].

4 Special ISI Case: Codewords Span Hyperplanes

In this section we analyze a special case of ISI. We assume that there are two codewords and that $K = N-1$. Whenever $K = N-1$ and $X_m (m = 1, 2)$ are full rank, $C_m (m = 1, 2)$ in (12) represent hyperplanes that pass through the origin. Thus we can find $\mathbf{p}^{(m)} \in \Re^N$ s.t.

$$C_m = \{\mathbf{x} \in \Re^N : \mathbf{x}^T\mathbf{p}^{(m)} = 0, \mathbf{p}^{(m)} \in \Re^N\}, m = 1, 2 \tag{20}$$

The distance between the hyperplane C_m and a vector $\mathbf{y} \in \Re^k$ is

$$d(\mathbf{y}, C_m) = \frac{|\mathbf{y}^T \mathbf{p}^{(m)}|}{||\mathbf{p}^{(m)}||} \tag{21}$$

The GLRT metric are given in $d(\mathbf{y}, C_m)$ by definition. The GLRT separating surfaces are therefore two hyperplanes given by:

$$s^G = \{\mathbf{v} : (\frac{\mathbf{p}^{(1)}}{||\mathbf{p}^{(1)}||} \pm \frac{\mathbf{p}^{(2)}}{||\mathbf{p}^{(2)}||})^T \mathbf{v} = 0\} \tag{22}$$

or equivalently

$$s^G = \{\mathbf{v} : \mathbf{p}^T \mathbf{v} = 0\} \tag{23}$$

where

$$\mathbf{p} = \frac{\mathbf{p}^{(1)}}{||\mathbf{p}^{(1)}||} \pm \frac{\mathbf{p}^{(2)}}{||\mathbf{p}^{(2)}||} \tag{24}$$

The normal to the surface at some $\mathbf{v} \in S^G$ intercepts C_1 at $\mathbf{a}^{(1)}$ and C_2 at $\mathbf{b}^{(2)}$. Fig. 1 illustrates a cross section of the hyperplanes for $N = 3$.

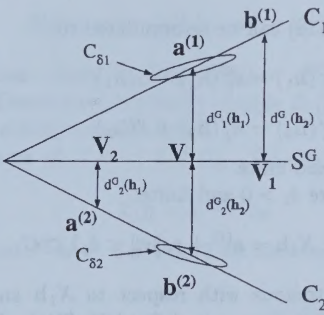


Figure 1: Cross Section for N=3

For X_1, X_2 full rank, we can find (unique) ISI parameters $\mathbf{h_1}$ and $\mathbf{h_2}$ s.t

$$\mathbf{a}^{(1)} = X_1 \mathbf{h_1} \tag{25}$$

and

$$\mathbf{b}^{(2)} = X_2 \mathbf{h_2} \tag{26}$$

where $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(2)}$ are defined above. According to definitions (13) and (14):

$$d_1^G(\mathbf{h_1}) = ||\mathbf{a}^{(1)} - \mathbf{v}|| \tag{27}$$

$$d_2^G(\mathbf{h_2}) = ||\mathbf{b}^{(2)} - \mathbf{v}|| \tag{28}$$

Define,

$$\mathbf{a}^{(2)} = X_2 \mathbf{h_1} \tag{29}$$

and

$$\mathbf{b}^{(1)} = X_1 \mathbf{h_2} \tag{30}$$

We can find $\mathbf{v_1}, \mathbf{v_2} \in S^G$ such that according to definitions (13) and (14):

$$d_2^G(\mathbf{h_1}) = \|\mathbf{a}^{(2)} - \mathbf{v_2}\| \tag{31}$$

$$d_1^G(\mathbf{h_2}) = \|\mathbf{b}^{(1)} - \mathbf{v_1}\| \tag{32}$$

We make the following assumption on the code:

$$d_1^G(\mathbf{h_1}) > d_2^G(\mathbf{h_1}) \tag{33a}$$

$$d_1^G(\mathbf{h_2}) > d_2^G(\mathbf{h_2}) \tag{33b}$$

where the inequalities are strict.

Note that we could have chosen, without loss of generality, the same assumptions with both signs of inequality reversed. If we cannot find any $\mathbf{v}$ such that these assumptions hold, we show in Section 5 that there is no decoder that uniformly improves the GLRT. We note, however, that we can easily find examples where the assumptions hold for some region of S^G. Under these assumptions we argue that there is a decoder that improves the order of the exponential error probability for $\mathbf{h_2}$ and either maintains or improves the exponential order of the error probability for the other possible channel coefficients. Follow the proof below.

The assumptions (33a) and (33b) can be reformulated to:

$$d_1^G(\mathbf{h_1}) = d_2^G(\mathbf{h_1}) + \alpha(\mathbf{h_1}) \tag{34a}$$

$$d_1^G(\mathbf{h_2}) = d_2^G(\mathbf{h_2}) + \beta(\mathbf{h_2}) \tag{34b}$$

where $\alpha(\mathbf{h_1}) > 0$ and $\beta(\mathbf{h_2}) > 0$ and finite.

Define the circle C_{δ_1} and where $\delta_1 > 0$ and finite:

$$C_{\delta_1} = \{\mathbf{h} : X_1\mathbf{h} = \mathbf{a}^{(1)} + \mathbf{r}, \|\mathbf{r}\| < \delta_1\} \subset C_1 \tag{35}$$

The distance function is continuous with respect to $X_1\mathbf{h}$ and $X_2\mathbf{h}$ and given by $d_m^G(\mathbf{h}) = |\mathbf{p}^{\mathrm{T}} X_m \mathbf{h}| / \|\mathbf{p}\|, m = 1, 2$ (where p is defined in (24)). We can therefore find δ_1 finite and small enough s.t.

$$\forall \mathbf{h} \in C_{\delta_1} \Rightarrow d_1^G(\mathbf{h}) = d_2^G(\mathbf{h}) + \alpha(\mathbf{h}), \alpha(\mathbf{h}) > 0 \tag{36}$$

Any new separating surface has to pass through some point on the line between $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(2)}$. The GLRT passes through $\mathbf{v}$. We look for a mapping from $\mathbf{v}$ to another point $\mathbf{u}$ for the new decoder. Under the assumptions (33a) and (33b) the new decoder we construct will pass at some point $\mathbf{u}$ between $\mathbf{v}$ and $\mathbf{a}^{(1)}$. In other words,

$$\mathbf{u} = \epsilon \mathbf{a}^{(1)} + (1 - \epsilon)\mathbf{v} \tag{37}$$

for some $0 < \epsilon < 1$. We first show that there exists $0 < \epsilon < 1$ small enough and finite such that the new surface will not damage the exponential error for all possible $\mathbf{h}$. That is, $\forall \mathbf{h}$ the vector $\mathbf{u}$ is strictly outside the balls of radius $\min\{d_1^G(\mathbf{h}), d_2^G(\mathbf{h})\}$ around $X_1\mathbf{h}$ and around $X_2\mathbf{h}$. Clearly, the point $\mathbf{u}$, which is in the decision region (of the GLRT decoder) of the codeword $m = 1$, is outside a ball of radius $\min\{d_1^G(\mathbf{h}), d_2^G(\mathbf{h})\}$ around $X_2\mathbf{h}$. It remains to show $\mathbf{u}$ is also outside a ball of radius $\min\{d_1^G(\mathbf{h}), d_2^G(\mathbf{h})\}$ around $X_1\mathbf{h}$.

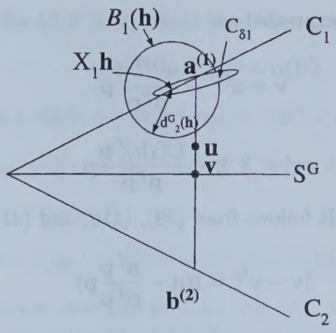


Figure 2: Illustration for $N = 3$ (Continued)

We split C_1 into two sets. The first set contains all $X_1\mathbf{h}$ s.t $\mathbf{h} \in C_{\delta_1}$. For any $\mathbf{h} \in C_{\delta_1}$ we define a ball of radius $d_2^G(\mathbf{h})$ around $X_1\mathbf{h}$:

$$\mathcal{B}_1(\mathbf{h}) = \{\mathbf{r} : \|X_1\mathbf{h} - \mathbf{r}\| < d_2^G(\mathbf{h})\} \tag{38}$$

See Fig. 2 for illustration. Since by (36) $d_2^G(\mathbf{h}) < d_1^G(\mathbf{h})$, the surface S^G is strictly separated from $\mathcal{B}_1(\mathbf{h})$. Therefore $\mathbf{v}$ is strictly outside $\mathcal{B}_1(\mathbf{h})$. It follows that there exists finite $\epsilon > 0$ s.t $\mathbf{u}$ is also strictly outside $\mathcal{B}_1(\mathbf{h})$.

The second set contains all $X_1\mathbf{h}$ s.t $\mathbf{h} \notin C_{\delta_1}$. We can therefore find finite $t > 0$ and unit vector $\mathbf{n}$ s.t.

$$X_1\mathbf{h} = \mathbf{a}^{(1)} + t\mathbf{n} \tag{39}$$

The projection of $X_1\mathbf{h}$ on S^G is $\mathbf{v}'$. See Fig. 3 for illustration.

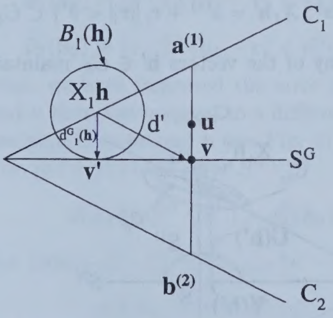


Figure 3: Illustration for $N = 3$ (Continued)

The distance of $X_1\mathbf{h}$ from S^G is

$$d_1^G(\mathbf{h}) = \|X_1\mathbf{h} - \mathbf{v}'\| \tag{40}$$

The vectors $\mathbf{v}$ and $\mathbf{v}'$ can be expressed as:

$$\mathbf{v} = \mathbf{a}^{(1)} - \frac{\mathbf{a}^{(1)T}\mathbf{p}}{\mathbf{p}^T\mathbf{p}}\mathbf{p} \tag{41a}$$

$$\mathbf{v}' = X_1\mathbf{h} - \frac{(X_1\mathbf{h})^T\mathbf{p}}{\mathbf{p}^T\mathbf{p}}\mathbf{p} \tag{41b}$$

where $\mathbf{p}$ was defined in (24). It follows from (39), (41a) and (41b) that:

$$\|\mathbf{v} - \mathbf{v}'\| = t(\mathbf{n} - \frac{\mathbf{n}^T\mathbf{p}}{\mathbf{p}^T\mathbf{p}}\mathbf{p}) \tag{42}$$

According to (39) the vector $\mathbf{n}$ is in the direction of $X_1\mathbf{h} - \mathbf{a}^{(1)}$ and according to (41a) the vector $\mathbf{p}$ is in the direction of $\mathbf{a}^{(1)} - \mathbf{v}$. The three vectors $\mathbf{a}^{(1)} \in C_1$, $X_1\mathbf{h} \in C_1$ and $\mathbf{v} \notin C_1$ are not all on the same line. Therefore $\mathbf{n} \neq \mathbf{p}/\|\mathbf{p}\|$ and $\mathbf{n} - (\mathbf{n}^T\mathbf{p})/(\mathbf{p}^T\mathbf{p})\mathbf{p} \neq 0$. It follows from (42) that since t is finite $\|\mathbf{v} - \mathbf{v}'\| > 0$ is also finite.

Denote by d' the distance of $X_1\mathbf{h}$ from $\mathbf{v}$. Since $\|\mathbf{v} - \mathbf{v}'\| > 0$ is finite it follows from the triangle inequality that there exists finite $\alpha(\mathbf{h}) > 0$ s.t.

$$d' = d_1^G(\mathbf{h}) + \alpha(\mathbf{h}) \tag{43}$$

For any $\mathbf{h} \notin C_{\delta_1}$ we define a ball of radius $d_1^G(\mathbf{h})$ around $X_1\mathbf{h}$:

$$\mathcal{B}_1(\mathbf{h}) = \{\mathbf{r} : \|X_1\mathbf{h} - \mathbf{r}\| < d_1^G(\mathbf{h})\} \tag{44}$$

Since $d' > d_1^G(\mathbf{h})$, the vector $\mathbf{v}$ is strictly outside $\mathcal{B}_1(\mathbf{h})$. Thus for all $\mathbf{h} \notin C_{\delta_1}$ there exists finite $\epsilon > 0$ s.t $\mathbf{u}$ is also strictly outside $\mathcal{B}_1(\mathbf{h})$.

We define now a circle of radius $\delta' < \delta_1$:

$$C_{\delta'} = \{\mathbf{h}' : X_1\mathbf{h}' = \mathbf{a}^{(1)} + \mathbf{r}, \|\mathbf{r}\| < \delta'\} \subset C_{\delta_1} \tag{45}$$

See Fig. 4 for illustration. Any of the vectors $\mathbf{h}' \in C_{\delta'}$ maintains the same conditions

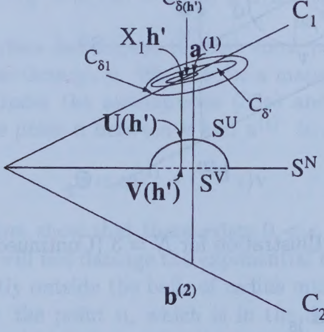


Figure 4: Illustration for $N = 3$ (Continued)

that $\mathbf{h_1}$ does. That is, for all $\mathbf{h}' \in C_{\delta'}$ there is $\alpha(\mathbf{h}') > 0$ s.t.

$$d_1^G(\mathbf{h}') = d_2^G(\mathbf{h}') + \alpha(\mathbf{h}') \tag{46}$$

and in addition there is a circle of a finite radius $\delta(\mathbf{h}') < \delta'$ around $X_1\mathbf{h}'$

$$C_{\delta(\mathbf{h}')} = \{\mathbf{h} : X_1\mathbf{h} = X_1\mathbf{h}' + \mathbf{r}, \|\mathbf{r}\| < \delta(\mathbf{h}')\} \subset C_{\delta'} \tag{47}$$

that also maintains:

$$\forall \mathbf{h} \in C_{\delta(\mathbf{h}')} \Rightarrow d_1^G(\mathbf{h}) = d_2^G(\mathbf{h}) + \alpha(\mathbf{h}), \alpha(\mathbf{h}) > 0 \tag{48}$$

The definition of $C_{\delta'}$ means that any $\mathbf{h}' \in C_{\delta'}$ not only maintains $d_1^G(\mathbf{h}') > d_2^G(\mathbf{h}')$ but also has a set around it that also maintains the same condition. The existence of the set $C_{\delta'}$ follows again from the continuity of the distance function and the fact $\alpha(\mathbf{h_1})$ is finite positive.

The projection vector of a certain $\mathbf{h}' \in C_{\delta'}$ on S^G is $\mathbf{v}(\mathbf{h}')$. Using the same arguments as for $\mathbf{h_1}$, the vector $\mathbf{v}(\mathbf{h}')$ can be mapped to $\mathbf{u}(\mathbf{h}')$:

$$\mathbf{u}(\mathbf{h}') = \epsilon(\mathbf{h}')X_1\mathbf{h}' + (1 - \epsilon(\mathbf{h}'))\mathbf{v}(\mathbf{h}') \tag{49}$$

for some $0 < \epsilon(\mathbf{h}') < 1$. The parameter $0 < \epsilon(\mathbf{h}') < 1$ is finite and will be chosen such that for any possible $\mathbf{h} \in \mathcal{R}^K$, the vector $\mathbf{u}(\mathbf{h}')$ will be strictly outside a ball of radius $\min\{d_1^G(\mathbf{h}), d_2^G(\mathbf{h})\}$ around $X_1\mathbf{h}$.

That is we have mapped an area of the separating surface without damaging the the exponential error probability for any possible channel vector.

Now it remains to show how the error probability for $\mathbf{h_2}$ has changed. First, we have not damaged the error probability for $\mathbf{h_2}$ since the new surface is strictly separated, by construction, from balls $\mathcal{B}_1(\mathbf{h_2})$ and $\mathcal{B}_2(\mathbf{h_2})$ of radius $\min\{d_1^G(\mathbf{h_2}), d_2^G(\mathbf{h_2})\} = d_2^G(\mathbf{h_2})$ around both $\mathbf{b}^{(2)} = X_2\mathbf{h_2}$ and $\mathbf{b}^{(1)} = X_1\mathbf{h_2}$:

$$\mathcal{B}_1(\mathbf{h_2}) = \{\mathbf{r} : \|X_1\mathbf{h_2} - \mathbf{r}\| < d_2^G(\mathbf{h_2})\} \tag{50}$$

$$\mathcal{B}_2(\mathbf{h_2}) = \{\mathbf{r} : \|X_2\mathbf{h_2} - \mathbf{r}\| < d_2^G(\mathbf{h_2})\} \tag{51}$$

We will show now that we have improved the error probability for $\mathbf{h_2}$. Denote by $S^V \subset S^G$ the area around $\mathbf{v}$ that was mapped to a different area (around $\mathbf{u}$). Denote by S^U the corresponding mapping area around $\mathbf{u}$ (see Fig. 4). The separating surface of the new decoder is defined by $(S^G \setminus S^V) \cup S^U$. For $S^G \setminus S^V$

$$d(X_2\mathbf{h_2}, S^G \setminus S^V) > d_2^G(\mathbf{h_2}) \tag{52}$$

according to construction (where $d(\mathbf{v}, S) = \min_{\mathbf{v}' \in S} \|\mathbf{v}' - \mathbf{v}\|$). For S^U

$$d(X_2\mathbf{h_2}, S^U) > d_2^G(\mathbf{h_2}) \tag{53}$$

since every $\mathbf{u}' \in S^U$ is in the decision area (of the GLRT decoder) of $m = 1$. In particular, there exists finite $\epsilon' > 0$ s.t.

$$\begin{aligned} d(X_2\mathbf{h_2}, \mathbf{u}) &= \|\epsilon\mathbf{a}^{(1)} + (1-\epsilon)\mathbf{v} - \mathbf{b}^{(2)}\| \\ &= \|\epsilon(\mathbf{a}^{(1)} - \mathbf{b}^{(2)}) + (1-\epsilon)(\mathbf{v} - \mathbf{b}^{(2)})\| \\ &= \|(\epsilon + \epsilon')(\mathbf{v} - \mathbf{b}^{(2)}) + (1-\epsilon)(\mathbf{v} - \mathbf{b}^{(2)})\| \\ &= (1+\epsilon')d_2^G(\mathbf{h_2}) > d_2^G(\mathbf{h_2}) \end{aligned} \tag{54}$$

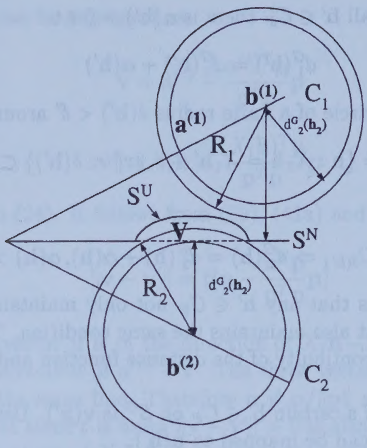


Figure 5: Illustration for $N = 3$ (Continued)

where the existence of ϵ' follows from the fact that $\mathbf{v}$ is on the line between $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(2)}$. Since $\mathbf{v}$ is a unique tangent point of S^G to $B_2(\mathbf{h_2})$ (uniqueness of projection on a hyperplane) and there is an area around $\mathbf{v}$ with finite radius that were also mapped to points with larger distance from $\mathbf{b}^{(2)}$, there is a ball around $\mathbf{b}^{(2)}$ with radius $R_2 > d_2^G(\mathbf{h_2})$ that is strictly separated from the new separating surface. See Fig. 5 for illustration. As for $X_1\mathbf{h_2}$,

$$d(X_1\mathbf{h_2}, S^N) > d^G(\mathbf{h_2}) \tag{55}$$

according to construction, and the inequality is strict. Therefore there is a ball of radius $d_2^G(\mathbf{h_2}) < R_1 \leq d_1^G(\mathbf{h_2})$ that is strictly separated from the new separating surface. Thus, we have managed to construct a surface with an improved minimal distance from $\mathbf{b}^{(2)}$ and thus improved the error probability for $\mathbf{h_2}$.

Note that the procedure can be repeated for any $\mathbf{v} \in S^G$ which maintains assumptions (33a) and (33b) (or with both inequality signs reversed) and the separating surface can be modified accordingly. Thus, the error probability can be improved for additional channel coefficients as well without damaging the error probability for any possible channel coefficient.

The decoding is done in the following way. Assume we receive the vector $\mathbf{y}$. The projection vector of $\mathbf{y}$ on S^G intersects S^G, C_1, C_2 at $\mathbf{v}, \mathbf{a}^{(1)}, \mathbf{b}^{(1)}$ respectively. Now if (33a) and (33b) hold and ϵ is given, we can find $\mathbf{u}$ by (37). If $\mathbf{y}$ is on the line between $\mathbf{a}^{(1)}$ and $\mathbf{u}$ we decode $\hat{m} = 1$ otherwise $\hat{m} = 2$. Note that this decoding rule depends on ϵ. It turns out, however, that the optimal ϵ is complicated to find. We currently look for a procedure that either determines an appropriate ϵ or does not require an explicit value of ϵ.

5 Converse Theorem

We prove now that the existence of $\mathbf{v} \in S^G$ such that both (33a) and (33b) hold (or both assumptions with inequality signs reversed) is also a necessary condition for the existence of a decoder that improves error probability uniformly (for which (19a) and (19b) hold).

Assume there is no $\mathbf{v} \in S^G$ such that both (33a) and (33b) hold (or both assumptions with inequality signs reversed). Therefore there are now only two possible cases for each $\mathbf{v} \in S^G$, in case I:

$$d_1^G(\mathbf{h_1}) < d_2^G(\mathbf{h_1}) \tag{56}$$
$$d_1^G(\mathbf{h_2}) > d_2^G(\mathbf{h_2}) \tag{57}$$

In case II,

$$d_1^G(\mathbf{h_1}) > d_2^G(\mathbf{h_1}) \tag{58}$$
$$d_1^G(\mathbf{h_2}) < d_2^G(\mathbf{h_2}) \tag{59}$$

Refer to case I. Any decoder is defined by a separating surface. Any separating surface has to pass at some point between $\mathbf{a}^{(1)}$ and $\mathbf{b}^{(2)}$. We claim that in case I in order to maintain (19a) the separating surface has to pass through $\mathbf{v}$. Therefore we were not able to achieve a smaller error probability for $\mathbf{h_1}$ (and $\mathbf{h_2}$). Refer now to case II. Considering $\mathbf{a}^{(2)}$ (defined in (29)) we project it on S^G at point $\mathbf{v}'$, and find the intersection of the difference vector $\mathbf{v}' - \mathbf{a}^{(2)}$ with C_1 at $\mathbf{d}^{(1)} = X_1\mathbf{h_3}$. Under our assumptions we have,

$$d_1^G(\mathbf{h_1}) > d_2^G(\mathbf{h_1}) \tag{60}$$
$$d_1^G(\mathbf{h_3}) < d_2^G(\mathbf{h_3}) \tag{61}$$

From the same arguments as in case I we cannot improve the error probability of $\mathbf{h}_1$ (and $\mathbf{h}_2$) since any separating surface maintaining (19a) has to pass through $\mathbf{v}'$. Since the above argument is valid for any $\mathbf{h}_1$ (and $\mathbf{h}_2$) the proof is complete.

6 Conclusions and Further Research

We were able to show that under certain conditions on the codewords the GLRT decoder for unknown ISI channels can be uniformly improved. This result can be directly applied to any unknown linear system, such as diversity channels with multiple antennas at the transmitter and a single antenna at the receiver.

We are also able to modify the above proof for the general case in which the codewords span subspaces and also to extend the result to M codewords. We are currently looking for an efficient procedure to implement the uniformly improved decoder.

References

[1] N. Seshardi V. Tarokh and A.R. Calderbank. Space-time codes for high data rate wireless communication: Performance criterion and code construction. *IEEE Trans. Inform. Theory*, 44(2):744–765, Mar. 1998.

[2] M. Feder and N. Merhav. Relative-minimax composite hypothsis testing and its applications. *IEEE Trans. Inform. Theory*, submitted for publication.

On Optimal Control over Packet-Oriented Communication Links*

Mikael Johansson and Lin Xiao

Information Systems Laboratory
Stanford University
Stanford CA 94305, USA
{mikaelj, lxiao}@stanford.edu

Abstract

Many advanced control systems use communication networks to send measurement and control signals between nodes. So far, the research on networked control system has been concerned exclusively with single data transfer. Today, however, transfer times are often a negligible part of the communication delay and many network protocols allow us to send several bytes of data at the same cost as a single-byte transfer. In this paper we explore how packet-oriented communication can be used to improve control system performance and robustness in presence of time-varying communication delays and link failures. We derive optimal controllers and estimators that exploit packet-oriented communication and show that they perform better than approaches based on single-byte transfer.

1 Introduction

In this paper we consider control of distributed systems where sensor data and actuator commands are sent over a communication network. This situation occurs frequently in complex systems where controllers need to exchange sensor data and control commands between spatially distributed system components. Networked control systems can since long be found in aircrafts, cars, and manufacturing systems. It is likely that more and more control loops will be closed over networks in the near future, as data networks become increasingly ubiquitous and prices of networks components continue to drop.

An understanding of the issues in networked control systems has just started to emerge and much work remains to be done. The theoretical problems associated with networked control systems depend on the nature of the communication. In systems that use media access based on time-division-multiplexing one faces problems on how to optimally schedule sensor measurement and control commands and how to design controllers for (periodically) time-varying systems (see *e.g.* [IF00]). In systems that use carrier-sense-multiple-access, on the other hand, nodes are allowed to transmit only when no other transmission (carrier) is present. This means that nodes might have to wait for the communication link to become available and need to resolve conflicts that occur when

---

*This research is supported in part by the Swedish Foundation for International Cooperation in Research and Higher Education, the Foundation Blanceflor-Boncompagni née Bildt, and DARPA under the Software Enabled Control Program (administered by AFRL under contract F33615-99-C-3014) .

several nodes try to transmit at the same time. From a theoretical point-of-view, this raises questions of how to chose priorities and how to design controllers for systems with time-varying distributed (sensing, actuation) delays, see *e.g.* [LR90, Nil98]. Different solutions are possible depending on whether the controller can determine the current delay or not, and whether the probability distributions of the communication delays are known or only hard bounds are available. Finally, some networks are unreliable in the sense that the communication link occasionally fails and packets are sometimes dropped.

The research on networked control systems has been exclusively devoted to single-byte data transfer. Today, however, transfer times are usually a negligible part of the communication delay and most network protocols allow us to send several bytes of data at no extra cost. This makes it interesting to investigate how multiple-byte transfers can be exploited to improve system performance and robustness to varying time delays and occasional loss of data. This paper details several uses of multiple-byte transfers.

The first idea that we investigate is to send a whole sequence of controls rather than a single piece of data. This allows both for sending predictive controls to be used in absence of new data, and to implement control laws that depend on the controller-actuator delay. As an illustration, we derive a optimal linear-quadratic regulator for a system with random independent communication delays and show how predictions and multiple-byte transfer give improved performance. The second idea that we investigate tries to mitigate the effects of a congested communication link between sensor and controller. More specifically, we consider a communication link with a non-zero probability that the sensor cannot communicate its measurement during the following sample interval. For this situation it has been suggested to use a try-once-discard protocol [WYB99] where data that cannot immediately be sent is discarded and only the most recent piece of data is sent. In a system that uses packet-oriented communication there is little need for discarding data, since the communication cost is the same irrespectively of the number of bytes sent. A simple idea is then to store data that cannot be sent in a buffer and transmit the full buffer content as a single packet once the communication channel becomes available. We investigate the optimal estimators under both information structures. Simulation studies indicate relatively moderate differences in long-term average performance of the two estimators and we give a simple explanation for this. In addition, we give scenarios where the significant improvements over the try-once-discard scheme are obtained.

2 Real-Time Control over Communication Networks

We consider a real-time control system where sensor and actuator data are transmitted over a communication network, as illustrated in Figure 1. Our focus will be on the single-loop case, although in general, the system might have multiple actuators and sensors connected through separate communication links.

There are two key ingredients that make control design for this type of systems challenging: (1) the dynamic effects of communication (primarily delay, but also possible link failures and lost packets) enter the feedback loop, and (2) the controller functionality (sensing, computation, and actuation) is implemented in a distributed fashion.

Different control schemes have been suggested for different variations of the set-up in Figure 1. For example, Luck and Ray [LR90] assume bounded communication delays and eliminate the delay variations by introducing timed buffers at the controller and actuator nodes. This simplifies the control design, since the system seen by the controller is now time-invariant, but results in longer-than-necessary delays that impair the achievable

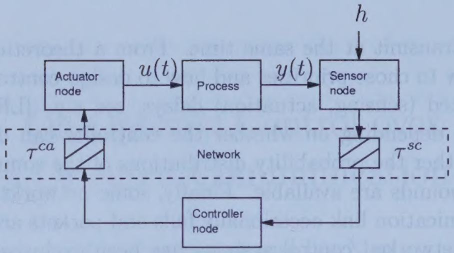


Figure 1: Networked control system.

performance. Nilsson [Nil98] derives optimal LQG controllers for a number of stochastic delay models assuming that the controller and actuator nodes are event-driven and carry out their operations as soon as new data arrives. It is assumed that the communication delay is less than one sample period and performance improvements over what can be achieved using timed buffers is reported. Xiao et.al [XHH00] assume that all nodes work synchronously and show how long delays, dropped packets, and various actuation policies can be conveniently dealt with in a jump-linear systems framework.

An important idea in networked control is the use of time-stamped messages. Time stamping simply means that the time-of-generation is appended to each message. This allows the receiver to determine the current communication delay by comparing the time-stamp with its local clock. As shown in [Nil98], time-stamping greatly simplifies the estimation problem, effectively rendering the estimation problem for a stochastic system into an estimation problem for a time-varying system, for which the standard Kalman filter is optimal. Unfortunately, this idea is not easily applied in the controller-actuator link, since it is assumed that the controller does all computations. This typically means that one has to resort to delay-independent control laws.

3 Exploiting Multiple-byte Transfer in Control

Today, many of the communication protocols that are used for networked control (Ethernet, Lonworks, Profibus, etc.) are not limited to single-byte data but support multiple-byte packets. Our first contribution is explore how this fact can be used to improve performance and robustness to link failures in the controller-actuator link.

3.1 Uses of Multiple-Byte Control Packets

Whether or not multiple-byte transfers can be exploited in the controller-actuator communication depends the computing capabilities of the actuator node. However, as we will see, even very modest computing capabilities (such as buffering and selection) enables several interesting uses of packet-oriented communication.

When the communication link is unreliable, or has long worst-case delay, it is natural to use multiple-byte transfers to send a sequence of predictive controls. Rather than computing a single command u_k, we then compute and transmit a sequence

$$\mu_k = \begin{bmatrix} u_k & \hat{u}_{k+1|k} & \dots & \hat{u}_{k+D|k} \end{bmatrix}$$

where D denotes the worst-case communication delay. The control strategies proposed in [Nil98, XHH00] hold the most recent control signal in absence of new data. In the next section, we will show that the use of predictions and multiple-byte transfer allow us to improve the performance of control strategies that use single-byte transfers.

Another useful idea is to send a sequence of *possible* controls, each optimal for a certain communication delay/latency in the controller-actuator link. The actuator node, which has full access to communication delays, then selects and applies the appropriate control from the packet. This allows us to implement control strategies that depend on the controller-actuator delay. For example, the modeling framework in [XHH00] allows systems with long communication delays and lost packets. If packets arrive in a different order than they were sent, the actuator policy is to discard the older of the two and always use controls from the most recent packet. However, the optimal control to transmit at time k depends on the latency[1] l_k of the packet currently in the actuator buffer

$$u_k = -L_k(l_k)x_k$$

which is unknown to the controller. Since the system in [XHH00] works synchronously, the latency takes on the values $\{0, h, \dots, Dh\}$. Hence, to implement the optimal control we can transmit

$$\mu_k = -\begin{bmatrix} L_k(0)x_k & L_k(h)x_k & L_k(2h)x_k & \dots & L_k(Dh)x_k \end{bmatrix}$$

and let the actuator select the appropriate control. To be able to do so, the actuator needs to store the D most recent latencies.

To illustrate the ideas, we will derive an LQ-optimal controller that uses predictions and multiple-byte transfers and show that it has superior performance to what is achieved using the control strategies in [LR90, Nil98].

3.2 An illustration: LQ-control incorporating predictions

Consider the system in Figure 1 and assume that the total communication delays satisfy

$$\tau_k \triangleq \tau_k^{sc} + \tau_k^{ca} \leq h$$

for all k. In [Nil98] an optimal linear-quadratic design procedure was derived for the following set-up: the sensor samples the process output with a regular sampling interval h while both controller and actuators are event-driven and carry out their operations as soon as new data is received. The actuator holds the control signal until the next piece of data arrives. The operation of the system is illustrated in Figure 2 (left), where the shaded regions illustrate the time intervals $[kh, kh+\tau_k)$ where the control u_{k-1} computed based on measurements at time $kh - h$ is held in absence of new data.

One drawback with this approach is the sometimes long hold times for the control (up to $2h$). The hold times can be reduced by use of predictive controls and multiple-byte transfers. When the controller receives the measurement sampled at time kh, we suggest to compute and transmit two commands; one that is applied immediately when the control packet arrives at the actuator, and one (predictive control) that is applied if no new command has arrived at the actuator the next sampling instant $kh + h$. The operation of this control strategy is shown in Figure 2 (right). Now, rather than holding the control u_k we apply a predictive control $\hat{u}_{k+1|k}$ in the interval $[kh + h, kh + h + \tau_k)$.

Assume that the underlying process is a linear-time invariant system

$$dx(t) = Ax(t)dt + Bu(t)dt + d\widetilde{w}(t) \tag{1}$$

---

[1] Recall that the latency of a packet is defined as its time of creation minus the current time.

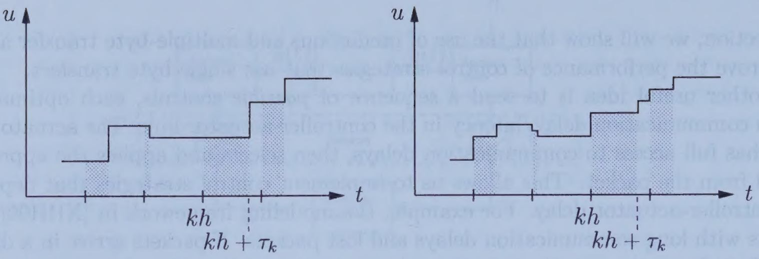


Figure 2: The asynchronous hold strategy (left) and predictive-corrective control (right).

where $\widetilde{w}(t)$ is Wiener process with zero mean and uncorrelated increments. Let the control objective be to minimize the continuous-time cost function

$$J = \mathbf{E}\left\{ \int_0^T x^T(t)Q_{1c}x(t) + 2x^T(t)Q_{12c}u(t) + u^T(t)Q_{2c}u(t)\, dt + x^T(T)Q_T x(T) \right\} \tag{2}$$

To keep things simple, we assume full and noise-free measurement of the state vector. Using the results in [ÅW97] we can convert this problem to an equivalent discrete-time LQR problem. The asynchronous hold strategy in [Nil98] can be modeled as

$$\begin{bmatrix} x_{k+1} \\ \hat{u}_k \end{bmatrix} = \begin{bmatrix} \Phi(h) & \Phi(h-\tau_k)\Gamma(\tau_k) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_k \\ \hat{u}_{k-1} \end{bmatrix} + \begin{bmatrix} \Gamma(h-\tau_k) \\ I \end{bmatrix} u_k + \begin{bmatrix} I \\ 0 \end{bmatrix} w(kh) \tag{3}$$

while the predictive-corrective control strategy suggested above is described by

$$\begin{bmatrix} x_{k+1} \\ \hat{u}_k \end{bmatrix} = \begin{bmatrix} \Phi(h) & \Phi(h-\tau_k)\Gamma(\tau_k) \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_k \\ \hat{u}_{k-1} \end{bmatrix} + \begin{bmatrix} \Gamma(h-\tau_k) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} u_k \\ \hat{u}_k \end{bmatrix} + \begin{bmatrix} I \\ 0 \end{bmatrix} w(kh). \tag{4}$$

For convenient notation, we introduce the augmented state and control vectors

$$z_k = \begin{bmatrix} x_k \\ \hat{u}_k \end{bmatrix}, \qquad \mu_k = \begin{bmatrix} u_k \\ \hat{u}_k \end{bmatrix}.$$

In both cases, the cost (2) can be transformed to an discrete-time equivalent

$$J(kh) = \mathbf{E}\left\{ \sum_{k=0}^{N-1} [z_k^T Q_1(\tau_k) z_k + 2 z_k^T Q_{12}(\tau_k)\mu_k + \mu_k^T Q_2(\tau_k)\mu_k] + z_N^T Q_T' z_N \right\} + N\overline{J} \tag{5}$$

where the term $N\overline{J}$ depends on the noise. If we assume that the communication delays are random and independent from transfer to transfer, the generalized dynamics under both control strategies can be described by a discrete-time stochastic system on the form

$$z_{k+1} = A(\tau_k)z_k + B(\tau_k)\mu_k + B_w(\tau_k)w_k \tag{6}$$

The optimal control can be derived using dynamic programming, similar to [Aok67]

Proposition 1 *Consider the system (6) with noise free measurement of the state vector* x_k *and assume that* $Prob(\tau_k > 0) \neq 0$. *The control law that minimizes the cost (5) is*

$$\mu_k = -L_k z_k$$

for $k = 0, 1, \dots, N-1$. The optimal feedback gains $\{L_k\}$ can be computed recursively, backward in time, by the following set of matrix difference equations

$$\widetilde{S}_k = \mathop{\boldsymbol{E}}_{\tau_k}\left\{\begin{bmatrix} Q_1(\tau_k) & Q_{12}(\tau_k) \\ Q_{12}^T(\tau_k) & Q_2(\tau_k)\end{bmatrix} + \begin{bmatrix} A^T(\tau_k) \\ B^T(\tau_k)\end{bmatrix} S_{k+1}\begin{bmatrix} A(\tau_k) & B(\tau_k)\end{bmatrix}\right\}$$

$$L_k = [\widetilde{S}_k]_{22}^{-1}[\widetilde{S}_k]_{21} \qquad S_k = [\widetilde{S}_k]_{11} - [\widetilde{S}_k]_{12}L_k$$

from the initial condition $S_N = Q'_T$.

Note that the asynchronous hold strategy derived in [Nil98] is a special case of the LQ-control proposed above in which the predictive and the corrective controls are identical. Hence, the predictive-corrective controller is guaranteed to perform better than the asynchronous hold strategy. To see if there are any noticeable performance gains in typical examples, we will evaluate the optimal state feedback controllers for both strategies on the example used in [Nil98, Section 5.5]. The process is on the form (1) with

$$A = \begin{bmatrix} 0 & 1 \\ -3 & -4\end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1\end{bmatrix}$$

and the incremental covariance of the state disturbance $\widetilde{w}$ is

$$R_{1c} = \begin{bmatrix} 35 \\ -61\end{bmatrix}\begin{bmatrix} 35 \\ -61\end{bmatrix}^T = \begin{bmatrix} 1225 & -2135 \\ -2135 & 3721\end{bmatrix}.$$

The matrices defining the cost function (2) are

$$Q_{1c} = 80\begin{bmatrix} \sqrt{35} \\ 1\end{bmatrix}\begin{bmatrix} \sqrt{35} \\ 1\end{bmatrix}^T, \qquad Q_{12c} = \begin{bmatrix} 0 \\ 0\end{bmatrix}, \qquad Q_{2c} = 1.$$

Note that we use the continuous-time cost criterion to enable proper comparison of the two control strategies. We will use a sample time of $h = 0.2$ (which is in line with the rule-of-thumb given in [ÅW97, Section 2.9]) and assume that the time delays are uniformly distributed on the interval $[0, h/2]$.

An upper bound on the performance is obtained by considering a system with a timed actuator buffer, offset by half a sample interval compared to the sensor. The optimal performance for this system is $J = 4.14 \cdot 10^4$. The optimal asynchronous hold controller achieves performance improvements of 15%, while the predictive-corrective controller improves the performance by 21%. Similar results were obtained in other examples.

The performance improvements of the asynchronous hold and the predictive-corrective control over the timed-buffer solution depend on the sampling interval. This is illustrated in Figure 3, where the relative performances of the asynchronous hold (rings) and the predictive-corrective controller (crosses) compared to the timed-buffer solution are shown for different values of the sampling period. As can be expected, the advantages of incorporating predictions are negligible for fast sampling (and hence fast communication) but increase as the sampling period and the range of possible input delays increase.

4 Exploiting Multiple-byte Transfer in Estimation

We will now investigate how packet-oriented communication can be exploited in the communication link between sensor and controller. To keep things simple, we assume that all units work synchronously and consider only the state estimation problem.

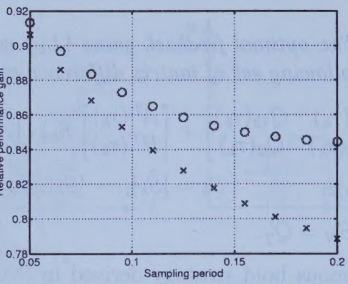


Figure 3: Relative performance gains as function of sampling interval.

Assume that we communicate the sensor readings over a data network with negligible transfer times but with a non-zero probability that the sensor cannot communicate its measurement during the coming sample interval. We let

$$\text{Prob(blocked)} = \alpha \qquad\qquad \text{Prob(send)} = 1 - \alpha.$$

For this type of systems, it has been suggested to use a try-once-discard scheme [WYB99] where information that cannot be sent immediately is discarded and only the most recent piece of information is sent. The intuition behind this protocol is that the latest sensor measurement contains the most information about the current state of the system. Moreover, refraining from sending data that is of little use results in an effective use of the communication link. In the packet-oriented framework, however, there is no need for discarding data since the communication cost is the same irrespectively of the size of the packet. A simple idea is then to store data that cannot be sent in a buffer. Once the sensor gets access to the communication link it sends the full buffer content in a single packet. We will refer to this as a store-and-send policy. Clearly, the number of packets sent using the two approaches is the same since we only send data when we get access to the communication link. We will derive the optimal estimators under both information structures, and compare their performances via Monte Carlo simulations.

Assume that the process is a linear discrete-time system on the form

$$x_{k+1} = Ax_k + Bu_k + B_w w_k \tag{7}$$

$$y_k = Cx_k + Du_k + D_v v_k. \tag{8}$$

where w_k and v_k are zero-mean Gaussian variables with covariance matrices R_w and R_v respectively. The standard Kalman filtering algorithm (valid also for time-varying systems) is given by

$$\hat{P}_{k|k-1} = A\hat{P}_{k-1}A^T + B_w R_w B_w^T \tag{9}$$

$$K_k = \hat{P}_{k|k-1}C^T(C\hat{P}_{k|k-1}C^T + D_v R_v D_v^T)^{-1} \tag{10}$$

$$\hat{P}_k = (I - K_k C)\hat{P}_{k|k-1} \tag{11}$$

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1} + Bu_{k-1} \tag{12}$$

$$\hat{x}_k = \hat{x}_{k|k-1} + K_k(y_k - \hat{x}_{k|k-1}) \tag{13}$$

If the sensor does not get access to the communication link at time k no data is transmitted to the controller node. This is equivalent to letting the linear system be time-varying, with $C = 0$, $D = 0$, $D_v = 0$ each time the communication link is congested. In terms of the Kalman filtering equations, this implies that if no data arrives at time k, $K_k = 0$ and

$$\hat{x}_k = \hat{x}_{k|k-1} \qquad\qquad \hat{P}_k = \hat{P}_{k|k-1}$$

In other words, we simply do an open loop propagation of the state and covariance estimates. This is the "suboptimal Kalman filter" suggested in [Che90], and it is also the optimal observer in a system that communicates using a try-once-discard protocol.

In a system with packet-oriented communication, on the other hand, we will receive all unsent data $\{y_k, y_{k+1}, \dots, y_K\}$, appropriately time-stamped, at some future time K. This allow us to re-compute the optimal state estimate $\hat{x}_K$ by re-iterating the Kalman filter equations from the optimal estimate $\hat{x}_{k-1}$ and the covariance estimate $\hat{P}_{k-1}$ using the recently received data $\{y_k, \dots, y_K\}$ and associated logged control signals $\{u_k, \dots, u_K\}$. This is the optimal estimator for a system that uses a store-and-send policy.

Before applying a try-once-discard protocol, it is important to be aware of the fact that observability can be lost under recurrent loss of data (see [LRH92]).

Example 1 (Loss of Observability using TOD Protocol) *Consider the system*

$$x_{k+1} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x_k, \qquad\qquad y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} x_k,$$

which is observable in the standard sense. Let the initial state be $x_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$. *This generates the output sequence*

$$\{y_k\} = \{1, 0, -1, 0, 1, 0, \dots\}.$$

Assume that every second sample is dropped by the try-once-discard protocol. The controller node might then receive the sequence (here "–" denotes that no data was received)

$$\{y_k\} = \{-, 0, -, 0, -, 0, \dots\}.$$

This sequence is indistinguishable from what would have been received if the initial state had been $x_0 = 0$, *and the system is unobservable. If all unsent data were transmitted when the sensor gets access to the communication link, the controller would receive the sequence*

$$\{y_k\} = \{-, (1, 0), -, (-1, 0), -, (1, 0), \dots\}$$

from which we could easily estimate the initial state.

Admittedly the example is somewhat contrived, but it illustrates that unfavorable traffic conditions can cause significant performance degradations if a try-once-discard protocol is used. To get a better understanding of typical estimator performance, we will compare estimator performances for random congestion sequences using Monte Carlo simulations. Assume that the underlying process is on the form (1) with

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -4 & -4 & -3 \end{bmatrix} \qquad\qquad B = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix}$$

and that the noise term $d\widetilde{w}(t)$ has unit incremental covaraiance. Let the sensor information $y_k = y(kh) = x_1(kh) + v_k$, with $v(k)$ being a zero-mean Gaussian random variable with $\mathbf{E}(v_k v_k^T) = 0.1$. We discretize the system using a sampling period of $h = 0.2$.

The mean-square errors for the two estimators as function of the congestion probability α are shown in Figure 4. In this example, the two estimators have relatively similar performance; not discarding data reduces the mean-square errors by less than 10%. The dashed curve shows the performance of the try-once-discard scheme using a time-invariant (and hence suboptimal) observer. The estimator uses predictions in the absence of new data but uses a constant gain matrix. The gains were chosen as the stationary Kalman filter gains for the process without delays in the measurements. The use of time-varying observers give large performance gains (up to 40%).

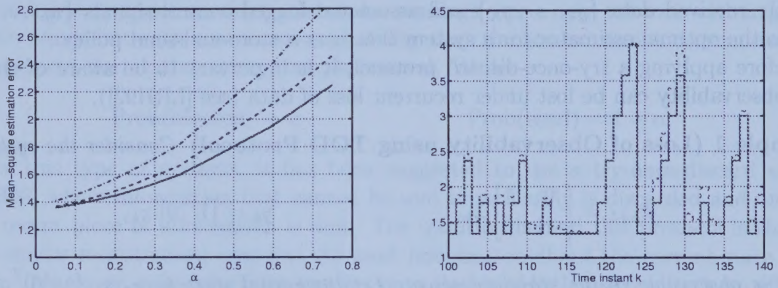


Figure 4: Mean-square errors as function of congestion probability α (left). Typical estimation error covariances for the different estimators (right).

To understand the performance curves better, note that the estimator works in two phases as illustrated in Figure 4 (right). In one phase, no measurement data is received and the estimation errors increase. In the other phase, recently recevied data is used to update the state and covariance estimates. For systems that use the store-and-send policy, these updates allow us to reset the state estimate and error covariance matrix to the values that would have been obtained under no data loss. The improvements are smaller when we use try-once-discard, since we only receive a single piece of data and only make one iteration of the Kalman filter equations. Notice, however, that the largest error reduction is obtained in the first iteration following a sequence of data losses, and that the average performance of the two filters are quite similar.

5 Conclusions

There has been a strong recent interest in analysis and design of networked control systems, where feedback loops are closed over a shared communication channel. The literature on networked control systems has been concerned exclusively with single data transfer. Today, however, transfer times are usually a small part of the total communication delay and many network protocol allows several bytes of data to be sent at no additional cost compared to single-byte data transfer. In this paper, we have explored a number of ideas for how multiple-byte transfers can be used to improve performance and robustness of networked control systems.

The first idea was to transmit a sequence of controls rather than a single command. This allows both for sending predictive controls to be used in absence of new data and to implement control laws that depend on the controller-actuator delay. To illustrate the ideas, we derived an optimal LQ controller that used predictions and showed that it always improves performance over the asynchronous hold strategies in [Nil98]. The second idea aimed at improving the performance of state estimators in the presence of a congested communication link between sensor and controller. We showed that observability can be lost if we use a protocol (TOD) that discards unsent data and only transmit the most recent sensor readings. We presented optimal observers under both TOD and packet-oriented communication, and showed improved long-term average performance in Monte Carlo simulations.

There are many possible extensions to the work presented in this paper. One interesting problem is how to design agile model predictive controllers for uncertain controller-actuator delay and possible packet loss. Another interesting question is how to take advantage of packet-oriented communication in systems that use time-division-multiplexing.

References

[Aok67] M. Aoki. *Optimization of stochastic systems: topics in discrete-time systems.* Academic Press, New York, 1967.

[ÅW97] K. J. Åström and B. Wittenmark. *Computer-Controlled Systems – Theory and Design.* Prentice Hall, third edition, 1997.

[Che90] G. Chen. A simple treatment for suboptimal Kalman filtering in case of measurement data missing. *IEEE Transactions on Aerospace and Electronic Systems*, 26(2):413–415, March 1990.

[IF00] H. Ishii and B. Francis. Stabilization with control networks. In *Control 2000*, Cambridge, UK, September 2000.

[LR90] R. Luck and A. Ray. An observer-based compensator for distributed delays. *Automatica*, 26(5):903–908, September 1990.

[LRH92] R. Luck, A. Ray, and Y. Halevi. Observability under recurrent loss of data. *J. Guidance Control and Dynamics*, 15(1):284–287, January/February 1992.

[Nil98] J. Nilsson. *Real-Time Control Systems with Delays.* PhD thesis, Department of Automatic Control, Lund Institute of Technologyy, Sweden, 1998. Also available from `http://www.control.lth.se/~johan`.

[WYB99] G. C. Walsh, H. Ye, and L. Bushnell. Stability analysis of networked control systems. In *Proceedings of the American Control Conference*, pages 2876–2880, San Diego, California, June 1999.

[XHH00] L. Xiao, A. Hassibi, and J. How. Control with random communication delays via a discrete-time jump system approach. In *Proceedings of the American Control Conference*, 2000.

Stochastic Transmission Policies for Time-Varying Channels with Memory Under Incomplete Channel State Information [*]

Danlu Zhang and Kimberly M. Wasserman
Electrical Engineering and Computer Science Department
University of Michigan
Ann Arbor, MI 48109
{danlu, wass@eecs.umich.edu }

Abstract

In this paper, we study the effect of channel memory, that is, memory in the error process, on the design of low energy data transmission protocols. We consider a single-user channel, and are interested in developing transmission schemes to determine when the user should transmit and at what power level so as to achieve a suitable balance between throughput and energy consumption. The channel state is not directly observable, and thus the transmission decisions must be based on the channel information, which may be partial or imperfect, provided by the base station over a feedback channel. We cast the problem as a partially observable Markov decision process, and provide the optimal transmission schemes for a number of different feedback scenarios; we also study the dependence of the structure of the optimal mechanisms on channel memory. Numerical results are provided.

1 Introduction

The presence of multipath fading and shadowing in wireless radio communication channels introduces memory in the form of statistical correlation between transmission errors. This memory can be exploited via adaptive transmission schemes employing channel prediction to appropriately adjust the operation of mobile users to the varying channel conditions.

For delay-tolerant applications, such as remote login, email, facsimile, and image and other file transfers, protocols at the data link and media access control (MAC) layers have the flexibility to defer transmissions and voluntarily release the channel if conditions deteriorate. Since each failed transmission is a waste of limited network resources, by making efficient control decisions, these adaptive transmission schemes can enhance the utilization of the channel as well as conserve energy for battery-operated mobile users.

Recent research has reexamined and modified traditional, persistent transmission/retransmission protocols in order to reduce the number of failed transmissions [1] [2] and [3]. In [1], a heuristic scheme is developed for deciding when to initiate and suspend transmission based on experimental data from a mobile user traveling at pedestrian speeds. A packet is transmitted only when the majority of a small window

---

[*] This work was supported in part by the National Science Foundation under grant ANI-9734025.

of past channel estimates indicates favorable channel conditions. This scheme results in a reduced number of failed transmissions at the expense of decreased throughput.

In [2] and [3] this tradeoff is interpreted as one between energy efficiency, defined as the ratio of the total amount of data delivered to the total amount of energy consumed, and throughput. They introduce a heuristic probing protocol, which suspends normal packet transmission upon failed transmissions, and enters a probing mode where small packets are periodically transmitted; when a probe packet is successfully acknowledged, normal packet transmission resumes. The performance of different automatic repeat request (ARQ) protocols is compared for varying degrees of correlation present in a Rayleigh fading process. They show that a conservative probing protocol performs well under slow fading conditions (low mobile velocities), while classic ARQ protocols yield better results when the fading process changes more rapidly (high mobile velocities).

In this paper, we are interested in (1) developing a control algorithm or transmission scheme for wireless data communication over channels with memory that determine when to attempt a transmission and at what power level, so as to achieve a suitable balance between throughput and energy consumption; and (2) studying the effect of channel memory on the performance, design, and structure of this scheme.

We assume the current state of the channel is not directly observable by the user, and that the transmission decisions must be based on the channel state information, which is partial or incomplete, provided to the user by the receiver over a feedback channel. We consider different feedback structures including: binary-only (*i.e.*, ack and nak) [14]; delayed channel state information provided at the end of those time slots during which packet transmissions are made; and delayed channel state information at the end of each slot, regardless of whether a transmission is made.

We approach the problem via stochastic control, namely, formulation of a partially observable Markov decision process (POMDP) [4]. A POMDP is an optimization model for sequential decision-making where multiple and conflicting objectives exist (*i.e.*, high throughput and high energy efficiency), there is uncertainty in the actual system state and outcomes of decisions, and present decisions impact future decisions. This approach provides a natural and systematic framework for the synthesis of transmission schemes in the presence of uncertainty, and tracking the structural dependence of the optimal transmission scheme with respect to channel memory. The conflicting goals of achieving high energy efficiency and throughput are reflected in a flexible cost structure that can be altered by either the mobile user or base station. The scheme that incurs the minimum total expected discounted cost over the infinite-horizon is obtained by applying techniques of stochastic dynamic programming.

A word on notation: Bold face letters denote vectors and matrices, and $\boldsymbol{a}'$ denotes the transpose of $\boldsymbol{a}$.

2 System model and problem formulation

The time-axis is divided into contiguous slots of equal duration which correspond to the time required to transmit a single fixed-length packet. The propagation and processing delays are taken to be negligible. To focus on the effect of channel memory, the buffer of the user is assumed to be continually backlogged with data packets that must be transmitted to the receiver. At the beginning of each time slot, the user must decide, based on the channel state information available at the time, whether or not to transmit a packet in the slot, and if so, at what power level. The algorithm or rule by which these decisions are made defines a transmission scheme.

2.1 Channel model, power control, and feedback structure

The uplink channel evolves as an M-state, irreducible, aperiodic, time-homogeneous Markov chain $G = \{G_n; n \in \mathbb{Z}_+\}$, where $G_n \in \mathcal{G} = \{g_1, g_2, \dots, g_M\}$ and $g_i \in (0, 1]$ is the state of the channel during the n^{th} time slot[1]. Assume $0 < g_1 < g_2 < \cdots < g_M$, with the interpretation that the greater the index, the better the channel quality. Let $\boldsymbol{R}$ be the probability transition matrix of G, with $(ij)^{th}$ element $R_{ij} = P\{G_n = g_j | G_{n-1} = g_i\}$, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_M)$ be the stationary distribution of G. Denote the i^{th} probability row-vector of $\boldsymbol{R}$ by $\boldsymbol{R}_i$. Let $\mathcal{P}_M$ be the set of all M-dimensional probability vectors.

For example, the Gilbert-Elliott (GE) channel ([8]) is a two-state channel with transition matrix

$$\boldsymbol{R} = \begin{bmatrix} 1-q & q \\ b & 1-b \end{bmatrix}. \tag{1}$$

The stationary distribution of $\boldsymbol{R}$ is given by $\pi_1 = b/(q+b)$ and $\pi_2 = q/(q+b)$.

To provide a definition of channel memory, consider the eigenvalues of the channel probability transition matrix. Recall from the Perron-Frobenius theorem, that the dominant eigenvalue of $\boldsymbol{R}$ is 1, and its corresponding (left) eigenvector can be normalized to yield the stationary distribution $\boldsymbol{\pi}$; the remaining $M-1$ eigenvalues have modulus less than 1. The initial distribution of G converges in total variation to $\boldsymbol{\pi}$ geometrically fast, with a rate bounded by the second dominant eigenvalue of $\boldsymbol{R}$ [11]. Based on this, we define the memory of a Markov channel as follows:

Definition 1 *The channel memory μ of G is defined to be the second dominant eigenvalue of the channel probability transition matrix $\boldsymbol{R}$.*

Note that for the GE channel in (1), second eigenvalue of $\boldsymbol{R}$ is given by $\mu = 1 - q - b$, which is defined to be the memory in [9]. The case of $\mu = 0$ corresponds to a memoryless channel, and $\mu = 1$ to a decomposable Markov chain, where the channel remains in its initial state for all time. In addition, for GE channels with non-negative memory ($\mu \geq 0$), the probability of remaining in a given state is greater than or equal to the stationary probability of being in that state, that is, for, say, state g_2 (the "good" state), $P\{G_n = g_2 | G_{n-1} = g_2\} = 1 - b \geq \frac{g}{g+b} = \pi_2$.

Recall that for two probability vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, $\boldsymbol{q}$ is said to be *stochastically less than* $\boldsymbol{p}$, denoted $\boldsymbol{q} \leq_{st} \boldsymbol{p}$, if and only if $\sum_{j=i}^{M} q_j \leq \sum_{j=i}^{M} p_j$, for all $i \in \{1, 2, \dots, M\}$ (see [10]). For two probability vectors on $\mathcal{G}$, if $\boldsymbol{q} \leq_{st} \boldsymbol{p}$, then $\boldsymbol{p}$ has more probability mass at the higher quality values of channel state than $\boldsymbol{q}$. For a multi-state channel, a natural generalization of non-negative memory is to assume the probability row-vectors of the channel probability transition matrix are stochastically increasing; quantitatively, $\boldsymbol{R}_{i-1} \leq_{st} \boldsymbol{R}_i$. Thus, the probability that the channel state is above any chosen level in the next time slot, given the current state is g_i is greater than if the current state is g_j, $j < i$. A probability transition matrix is said to be *monotone* if its probability row-vectors are stochastically increasing. Such matrices satisfy the following properties:

Theorem 1 ([12]) *If $\boldsymbol{R}$ is a monotone probability transition matrix, then $\boldsymbol{q}\boldsymbol{R} \leq_{st} \boldsymbol{p}\boldsymbol{R}$ for all probability distributions $\boldsymbol{q} \leq_{st} \boldsymbol{p}$; $\boldsymbol{R}^k$ is monotone; if $\boldsymbol{R}$ is irreducible, it has, apart from its simple eigenvalue at 1, a second non-negative eigenvalue at $\mu < 1$ such that all other eigenvalues have modulus not exceeding μ.*

We make the following assumption on the channel:

---

[1] For more information and details on Markovian models for wireless communication channels, see [5] [6] [13] and the references therein.

Assumption 1 *The channel probability transition matrix $\boldsymbol{R}$ of G is a monotone matrix.*

Since $\boldsymbol{R}$ is assumed monotone, $\mu \geq 0$ by Theorem 1. Note that for the GE channel in (1), the second row-vector is stochastically greater than the first, if $\mu \geq 0$.

Define $S_n \in \{0, u_1, u_2, \ldots, u_L\}$ to be the transmission power of the user during the n^{th} time slot, where $S_n = s > 0$ corresponds to transmitting a packet with power s watts, and $S_n = 0$ corresponds to suspending transmission (zero power). Let $\mathcal{S} = \{u_1, u_2, \ldots, u_L\}$ and $\mathcal{S}_0 = \mathcal{S} \cup \{0\}$. Define $P_e : \mathbb{R}_+ \rightarrow [0, 1]$ to be the probability of a packet error, which is a non-decreasing function of the received energy-to-interference density ratio, which is in turn proportional to the product of the channel state and the transmission power (*i.e.*, the received signal power). The functional form of P_e depends on parameters such as the modulation and channel coding schemes, interleaving depth, and the choice of receiver architecture.

At the end of each slot, the user obtains error-free feedback from the receiver specifying whether a packet was successfully received, and possibly providing outdated or delayed measurements of the channel state. Let $A_n \in \mathcal{A}$ be the feedback or observation available to the user at the beginning of the n^{th} time slot, where $\mathcal{A}$ depends on the feedback structure. We consider three scenarios of feedback structure:

(i) Binary feedback: After a packet transmission, the user is notified as to whether the packet was correctly received. In this case, $A_n = 0$ indicates a successful transmission during the $(n-1)^{st}$ slot, and $A_n = 1$ indicates either a failed transmission or a confirmation that a packet was not sent ($S_{n-1} = 0$); thus $\mathcal{A} = \{0, 1\}$. No information regarding the channel state is provided to the user; the channel is a hidden Markov model.

(ii) Delayed measurements at completions of transmissions: The state of the channel during a slot is made available to the user at the beginning of the next slot, only if a packet transmission was attempted by the user during that slot. In this case,

$$A_n = \begin{cases} G_{n-1} & \text{if } S_{n-1} > 0 \\ 1 & \text{if } S_{n-1} = 0, \end{cases}$$

where 1 indicates a confirmation that a packet was not sent; thus $\mathcal{A} = \{1, g_1, g_2, \ldots, g_M\}$. For $S_{n-1} > 0$, the user is also informed as to whether the packet transmitted in the $(n-1)^{st}$ slot was correctly received, though this is suppressed from the notation for simplicity.

(iii) Delayed measurements at the end of each slot: The state of the channel during a slot is made available to the user at the beginning of the next slot, regardless of whether a transmission is made in the slot. Thus $A_n = G_{n-1}$, and $\mathcal{A} = \{g_1, g_2, \ldots, g_M\}$. As above, for $S_{n-1} > 0$ the user is also informed as to whether the packet transmitted in the $(n-1)^{st}$ slot was received correctly, but again this is suppressed from the notation.

2.2 Partially observable Markov decision process formulation

The current state of the channel cannot be observed. An appropriate definition of state, called *information state* using the terminology in [4], at time n is then the conditional probability distribution $\boldsymbol{p}_n$ of the channel state G_n, given the initial probability distribution $\boldsymbol{p}_0$ on $\mathcal{G}$, the past transmission decisions $S_0, S_1, \ldots, S_{n-1}$, and the past and present observations $A_0, A_1, \ldots, A_n$:

$$p_n(i) = P\{G_n = g_i | \mathbf{p}_0 = \mathbf{p}; S_{n-1} = s_{n-1}, \cdots, S_0 = s_0; A_n = a_n, \cdots, A_0 = a_0\}, \quad (2)$$

where $i = 1, 2, \ldots, M$. Note $\boldsymbol{p}_n \in \mathcal{P}_M$ and thus the information state has $M - 1$ degrees of freedom.

If the channel is in state g and power $s \geq 0$ is used, a cost of $C(g, s)$ is incurred. An example cost assignment is given by

$$C(g, s) = \begin{cases} c_0 & s = 0 \\ c_1 s + c_2 P_e(gs) & s > 0. \end{cases} \tag{3}$$

It is clear that this cost assignment penalizes both overly conservative and aggressive transmission schemes. For fixed s, assume $C(g, s)$ is decreasing in g; and for fixed g, assume $C(g, s)$ is continuous in s and $\arg\min_{s\in\mathcal{S}}\{C(g_i, s)\} < \arg\min_{s\in\mathcal{S}}\{C(g_j, s)\}$, for $i < j$, so that the transmission powers optimizing the costs are decreasing in channel gain. Let $C(g_i, 0) = c_0$, $i = 1, 2, \ldots, M$, so the cost of suspending transmission does not depend on the channel state. Also, to avoid trivialities, assume $\min_{s\in\mathcal{S}}\{C(g_M, s)\} < c_0$, so the cost of suspending transmission when the channel is in the best state is greater than that of packet transmission.

Let $c(\boldsymbol{p}, s) \triangleq E[C(G_n, s)|\boldsymbol{p}_n = \boldsymbol{p}] = \sum_{i=1}^{M} C(g_i, s)p(i)$ be the immediate expected cost incurred at a time slot if the information state is $\boldsymbol{p}$ and power $s \geq 0$ is used. Note that $c_0 = c(\boldsymbol{p}, 0)$, $\boldsymbol{p} \in \mathcal{P}_M$, is the immediate expected cost of suspending transmission. Also since for fixed s, $C(g, s)$ is decreasing in g, if $\boldsymbol{q} \leq_{st} \boldsymbol{p}$ then $c(\boldsymbol{q}, s) \geq c(\boldsymbol{p}, s)$. This and Theorem 1 imply that if $g_j < g_i$ then $\boldsymbol{R}_j^m \leq_{st} \boldsymbol{R}_i^m$ and thus $c(\boldsymbol{R}_j^m, s) \geq c(\boldsymbol{R}_i^m, s)$. Finally, assume

$$\min_{s\in\mathcal{S}} c(\boldsymbol{\pi}, s) < c_0, \tag{4}$$

so that if the information state is the stationary distribution of G, the immediate expected cost of transmitting a packet is less than that of suspending transmission.

For notational convenience, let $c^*(\boldsymbol{p}) = \min_{s\in\mathcal{S}}\{c(\boldsymbol{p}, s)\}$, and $s^*(\boldsymbol{p}) = \arg\min_{s\in\mathcal{S}}^{-1}\{c(\boldsymbol{p}, s)\}$. We assume that $s^*(\boldsymbol{p})$ is concave in $\boldsymbol{p}$.

The above system describes a partially observable Markov decision process (MDP) (see [7]). Note that partially observable MDPs are generally more difficult to analyze than their widely used completely observable counterparts [7]. This is due to the fact that the state space of a partially observable MDP is often much larger than than a completely observable MDP, and thus solution techniques such as successive approximation and policy iteration are often much more computationally demanding. In addition, qualitative or structural properties for completely observable MDPs, such as threshold rules, often do not generalize to, or are more difficult to establish in, partially observable models. However, due to the special structure of our model, we are able to obtain the optimal solution.

We seek a transmission scheme that minimizes the expected total discounted cost

$$V(\boldsymbol{p}) = E\left[\sum_{n=0}^{\infty} \alpha^n C(G_n, s_n) \middle| \boldsymbol{p}_0 = \boldsymbol{p}\right], \tag{5}$$

where $\boldsymbol{p}$ is an initial probability distribution on $\mathcal{G}$ and $\alpha \in (0, 1)$ is a discount factor weighing present costs more heavily than future costs. The function $V : \mathcal{P} \to \mathbb{R}$ is called the value function.

It can readily be shown that to optimize (5) it is sufficient to base the transmission decision at time slot n on $\boldsymbol{p}_n$(see [7]). Also, for any fixed sequence of powers $\{s_n; n \geq 0\}$ and observations $\{a_n; n \geq 1\}$, the sequence of probabilities $\{\boldsymbol{p}_n; n \geq 0\}$ forms a Markov chain over the uncountable state space $\mathcal{P}$ (see [7]).

In the next section, we will make use of stochastic dynamic programming techniques to obtain the optimal transmission scheme, and solve for the value function. Note that the optimal transmission scheme is stationary, that is, time-invariant.

3 Binary feedback

The result for this scenario has been presented in [14] and is only outlined here. We restrict to the case where $\mathcal{S} = \{s\}$, $Pe(1) = Pe, 0 < Pe < 1$ and $Pe(2) = 0$. The optimal value function satisfies the following recursion (see [7]):

$$V(p) = \min\{c_0 + \alpha V(q + \mu p), \\ c(p, s) + \alpha[1 - P_e(1 - p)]V\left(q + \frac{\mu p}{1 - P_e(1 - p)}\right) + \alpha P_e(1 - p)V(q)\Big\}. \quad (6)$$

Here the information state only has one degree of freedom and we denote it as $p = p(2)$. We proceed by considering the following threshold-based transmission mechanism: There exists $p^* \in (q, \pi_2)$ such that the optimal transmission mechanism transmits a packet in a slot with power s if and only if $p > p^*$. We then obtain the value function and verify it satisfies (6) for optimality. We have shown that this threshold rule is optimal for a wide range of parameter values: $0 \leq \alpha \leq 1-10^{-6}, 0 \leq P_e \leq 0.95, 0 \leq \mu \leq 1-10^{-6}$, $1 \leq q/b \leq 1000$.

The above threshold rule can be interpreted as a simple backoff rule. Since $\mu \geq 0$, if the user transmitted during a slot and succeeded, $p = 1 - b \geq p^*$, and thus the user transmits in the next slot. If the transmission failed, $p = q \leq p^*$, and the user suspends transmission. It requires a finite number T_1 of time slots for the information state to exceed the threshold, where $T_1 = \left\lfloor \ln(\frac{\pi_2 - p^*}{\pi_2 - q})/\ln \mu \right\rfloor$.

4 Optimal transmission scheme of Scenario 2

The update equation for the information state at time n is thus given by

$$\boldsymbol{p}_n = \begin{cases} \boldsymbol{R}_i; & S_{n-1} > 0, A_n = G_{n-1} = g_i \\ \boldsymbol{p}_{n-1}\boldsymbol{R}; & S_{n-1} = 0. \end{cases} \quad (7)$$

The transmission decisions effect the future values of the feedback or observation, and hence the channel information to the user through the information state. Note that (7) depends on the decision of whether or not to transmit a packet, but not, if the decision is to transmit, on the choice of transmission power. The optimal value function satisfies the following recursion (see [7]):

$$V(\boldsymbol{p}) = \min\{c_0 + \alpha V(\boldsymbol{pR}), c^*(\boldsymbol{p}) + \alpha \sum_{j=1}^{M} V(\boldsymbol{R}_j)p(j)\}. \quad (8)$$

We will obtain the optimal transmission scheme by considering a finite horizon optimization problem, and then applying the method of successive approximation (see [4]). All the derivations are included in [15].

First, it can been shown that $V(\boldsymbol{p})$ is continuous and a concave function of $\boldsymbol{p}$, and if $\boldsymbol{q} \leq_{st} \boldsymbol{p}$ then $V(\boldsymbol{q}) \geq V(\boldsymbol{p})$. Note that for multi-state channels, not all $\boldsymbol{p} \in \mathcal{P}_M$ can be stochastically ordered, so it may be difficult to determine the optimal partitioning of $\mathcal{P}_M$ into regions of transmission and suspension. However, the following can be shown from which the structure of the optimal transmission scheme for multi-state channels and multiple power levels can be characterized.

Proposition 1 *Let the information state at the beginning of a time slot be* $\boldsymbol{p}$. *If* $c^*(\boldsymbol{p}) \leq c_0$ *then a packet transmission should be made in the slot with power* $s^*(\boldsymbol{p})$.

Thus Proposition 1 provides a sufficient, easy-to-verify condition for making a packet transmission in a slot. It is worth noting that the condition in Proposition 1 is not necessary for transmission since transmitting can also obtain a better observation of the channel. However, Proposition 1 is enough to characterize the structure of the optimal transmission scheme. In particular, at the end of a time slot during which a transmission was made, if the cost of transmitting given the current information state is less than the cost of suspending transmission, then transmission continues with a suitably chosen power; otherwise, transmission may be suspended for a finite number of slots depending on the state of the channel during the previous slot. Thus we have:

Proposition 2 *Fix a time slot and suppose a packet transmission occurred in the previous slot, during which the channel state was* g_i. *If* $c^*(\boldsymbol{R}_i) > c_0$, *then there exists* $0 \leq T_i < \infty$ *such that transmission is suspended for* T_i *consecutive time slots, and then resumes with power* $s^*(\boldsymbol{R}_i\boldsymbol{R}^{T_i})$.

For any given time slot and arbitrary value of the information state, by Proposition 2, the user will eventually transmit a packet. At the end of the time slot during which this transmission occurred, the user learns the channel state during that slot, and thus the information state at the beginning of the next slot is one of the row vectors of $\boldsymbol{R}$. Assuming the initial information state is one of the probability row vectors of $\boldsymbol{R}$, by Proposition 3, the state space of the information state is then given by the finite set $\{\boldsymbol{R}_i, \boldsymbol{R}_i\boldsymbol{R}, \boldsymbol{R}_i\boldsymbol{R}^2, \ldots, \boldsymbol{R}_i\boldsymbol{R}^{T_i}; i = 1, 2, \ldots, M\}$, which has cardinality $\sum_{i=1}^{M}(T_i + 1)$.

Since the user learns the actual channel state at the end of those slots during which it transmits, and the optimal scheme is a backoff rule, the user does not need to actually compute the information state (2) at the beginning of each time slot. Rather, for each probability row vector of $\boldsymbol{R}$, that is, channel state, there corresponds a backoff period, which is zero for at least one state, and transmission power, which is the power level used at the conclusion of the backoff period. At the end of a packet transmission, the user learns the channel state during that time slot; it then suspends transmission for the corresponding backoff period, after which it resumes transmission with the corresponding power. At the end of this transmission, the user again learns the actual channel state during that time slot. Note that there are only $M + 1$ possible transmission powers used (including zero).

Due to the structure of the optimal transmission scheme, the solution to the recursion (8) with initial state $\boldsymbol{R}_i$ can be found by solving the following set of linear equations:

$$V(\boldsymbol{R}_i) = \frac{c_0(1-\alpha^{T_i})}{1-\alpha} + \alpha^{T_i}c^*(\boldsymbol{R}_i\boldsymbol{R}^{T_i}) + \alpha^{T_i+1}\sum_{j=1}^{M} V(\boldsymbol{R}_j)R_{ij}^{T_i+1}, \tag{9}$$

where $R_{ij}^{T_i+1}$ is the $(ij)^{th}$ element of the matrix $\boldsymbol{R}^{T_i+1}$. The first term in (9) accounts for the cost of suspending transmission for T_i consecutive slots, and the second term for the future costs of transmitting on the $(T_i + 1)^{st}$ slot. Eventually, the expected total cost is then given by

$$V = \sum_{i=1}^{M} V(\boldsymbol{R}_i)\pi_i, \tag{10}$$

assuming the initial information state is chosen according to the stationary distribution of G. The optimal values of the backoff periods are those values of $\{T_i\}$ that minimize (10). We have the analytical formula for the Gilbert-Elliott channel with multiple power levels.

4.1 Scenario 3

This is the easiest scenario among all the three. The value recursion is given by

$$V(p) = \min\{c_0, c^*(p)\} + \alpha \sum_{j=1}^{M} V(R_j)p(j), \tag{11}$$

$p \in \{R_1, R_2, \ldots, R_M\}$. Thus the optimal transmission mechanism is a simple myopic rule: Given the current information state p, transmit with power $s^*(p)$ in the current slot if $c^*(p) \leq c_0$, and suspend transmission (zero power) otherwise. This can also be interpreted as a backoff rule.

5 Numerical results

Due to space limitations, we only present an example for Scenario 2. For convenience, we study channel gain in terms of the received SIR. We assume the short-term fading is averaged out over a slot, and track the long-term fading. We model the channel gain (SIR) as a log-normal random process with mean 8 dB (the average SIR when the normalized transmission power is one) and autocorrelation $R(n) = \sigma^2(0.3)^{\gamma|n|}$, where $\sigma^2 = 4.3$ dB and γ is proportional to the velocity of the user (see [13]). This implies that the channel can be modeled as a Markov process. We then quantize the channel to obtain a finite-state Markov chain, where the quantization intervals are chosen to make each state equiprobable. The transition probabilities are determined by suitably integrating over the conditional probability density function. For simplicity, we assume an uncoded system using BPSK modulation with 256-bit packets (codewords). We use the cost function in (3), with $c_0 = 1.0$, $c_1 = 0.09$, and $c_2 = 3.0$, and a peak transmission power of 4. We present results for 2- and 4-state channels.

In both cases, the state in which the user backs off is only the first (and worst) state. In Figures 1 and 2, we plot the optimal backoff times as a function of the channel memory for the 2- and 4-state channels, respectively. In each case, it can be seen that the optimal backoff time increases as the memory increases. This is due to the fact that the greater the memory, the longer the sojourn time in the current state.

In Figures 3 and 4, we plot the optimal transmission power upon conclusion of a backoff period for the 2- and 4-state channels, respectively. In Figure 3, if the channel was in the good state in the most recent time slot, then the backoff time $T_2 = 0$, and the optimal transmission power decreases with memory; otherwise, if the channel was in the bad state in the most recent time slot, then the backoff time $T_1 \geq 0$, and the optimal transmission power increases with memory. The jumps in Figure 3 correspond the jumps in the optimal backoff times in Figure 1. Similar behavior is exhibited in the 4-state case in Figure 4, though reasoning this analytically is more difficult.

In Figures 5 through 10, we plot the total expected discounted cost (since this is the fundamental quantity we are optimizing), throughput, and long-term average energy efficiency for the optimal transmission schemes, and compare to a classic transmission scheme which always transmits (using the optimal chosen power given the information state). The energy efficiency is defined to be the same as in [2], namely, the average throughput in steady-state divided by the average power consumption in steady-state. The performance with a 4-state model is better than with a 2-state model, which is due to a finer quantization of the channel. Though we don't present the results here, the difference in performance between a 4- and appropriate 6-state model was not drastic. Note that by appropriately selecting the cost assignments, the throughput and energy efficiency of the optimal transmission scheme can be chosen to meet various quality of service requirements, and still provide an energy savings, especially when the channel memory is high.

6 Conclusions

An analytical, stochastic control framework for designing transmission mechanisms for wireless data communication that are sensitive to the conflicting goals of high throughput and low energy operation has been presented. Optimal transmission scheduling and power control schemes have been provided for a variety of different feedback scenarios. All these optimal schemes can be interpreted as backoff rules and have simple implementation. The impact of the channel memory has also been studied.

Our future work includes extending the current result to the cases with errors in channel measurements, infrequent channel feedback and with multiple users.

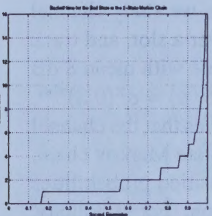


Figure 1: Backoff period for the bad state in the 2-state channel as a function of memory.

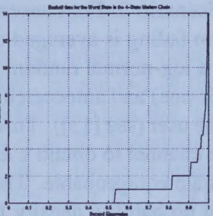


Figure 2: Backoff period for the worst state in the 4-state channel as a function of memory.

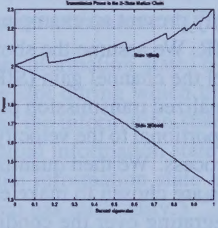


Figure 3: Transmission power at conclusion of backoff period for the 2-state channel as a function of memory.

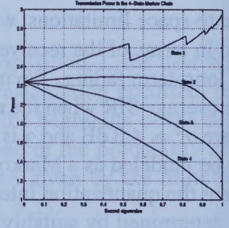


Figure 4: Transmission power at conclusion of backoff period for the 4-state channel as function of memory.

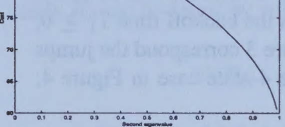


Figure 5: Cost for the 2-state channel as a function of memory.

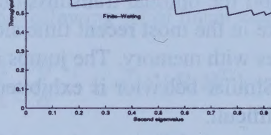


Figure 6: Throughput for the 2-state channel as a function of memory.

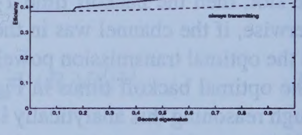


Figure 7: Energy efficiency for the 2-state channel as a function of memory.

References

[1] B. R. Badrinath and P. Sudame, "To send or not to send: Implementing deferred transmissions in a mobile host," in *Proc. 16th IEEE Internat. Conf. Distr. Comput. Sys.*, 1996.

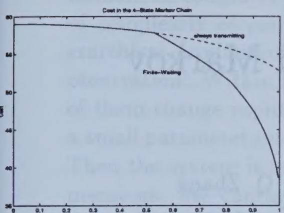


Figure 8: Cost for the 4-state channel as a function of memory.

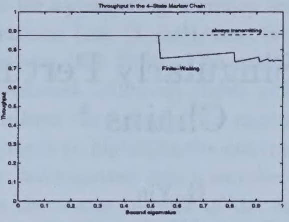


Figure 9: Throughput for the 4-state channel as a function of memory.

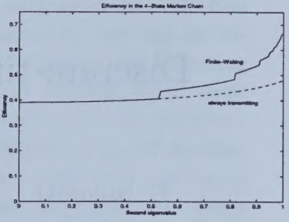


Figure 10: Energy efficiency for the 2-state channel as a function of memory.

[2] M. Zorzi and R. R. Rao, "Error control and energy consumptionfor nomadic computing," *IEEE Trans. Computer*, vol. 46, no. 3, pp. 279–289, 1997.

[3] M. Zorzi, R. R. Rao, and L. B. Milstein, "ARQ error control for fading mobile radio channels," *IEEE Trans. Veh. Tech.*, vol. 46, pp. 445–455, 1997.

[4] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[5] L. N. Kanal and A. R. K. Sastry. Models for channels with memory and their applications to error control. *Proc. IEEE*, vol. 66, no. 7, pp. 724–744, 1978.

[6] H. S. Wang and N. Moayeri. Finite-state Markov channel: A useful model for radio communication channels. *IEEE Trans. Veh. Tech.*, vol. 44, pp. 163–171, 1995.

[7] G. E. Monahan, "A survey of partially observable Markov decision processes: Theory, models, and algorithms," *Management Sci.*, vol. 28, no. 1, pp. 1–16, 1982.

[8] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Tech. J.*, vol. 39, pp. 1253–1266, 1960.

[9] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliott channels" *IEEE Trans. Inform. Theory*, vol. 35, no. 6, pp. 1277–1290, 1989.

[10] S. Ross, *Stochastic Processes*, 2/e. John Wiley & Sons, Inc., New York, NY, 1996.

[11] J. S. Rosenthal, "Convergence rates for Markov chains," *SIAM Review*, vol. 37, no. 3, pp. 387–405, 1995.

[12] J. Keilson and A. Kester, "Monotone matrices and monotone Markov processes," *Stoch. Proc. App.*, vol. 5, pp. 231–241, 1977.

[13] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *Elec. Letters*, vol. 27, no. 23, pp. 2145–2146, 1991.

[14] D. Zhang and K. M. Wasserman, "Energy efficient data communication over fading channels," Proc. IEEE Wireless Commun. and Networking Conf. (WCNC'00), Chicago, IL, 2000.

[15] D. Zhang and K. M. Wasserman, "Stochastic power control for wireless data communication over channels with memory", in preparation for submission to *IEEE/ACM Trans. Networking*.

Discrete-time Singularly Perturbed Markov Chains *

G. Badowski
Dept. of Mathematics
Wayne State University
Detroit, MI 48202
grazyna@math.wayne.edu

G. Yin
Dept. of Mathematics
Wayne State University
Detroit, MI 48202
gyin@math.wayne.edu

Q. Zhang
Dept. of Mathematics
University of Georgia
Athens, GA 30602
qingz@control.math.uga.edu

Abstract

Using singularly perturbed Markov chains to model stochastic network has gained much popularity recently. The motivation stems from various applications in manufacturing and production planning, stochastic network, and communication systems. Such systems are inevitable large dimensional and having many states. One uses singularly perturbed Markov chains to reflect different rates of change in the systems. The essence is to take advantage of the time-scale separation and to reduce the complexity of the underlying systems. Various asymptotic properties of the underlying Markov chains are studied including asymptotic expansion of the probability distribution vector, unscaled and scaled occupation measures, limit switching diffusions, and associated nearly optimal control problems.

Key words. Markov chain, discrete time, singular perturbation, switching diffusion, aggregation, weak convergence.

1 Introduction

Various applications in stochastic networks such as communication networks, wireless and mobile systems, and manufacturing flowshops, require using random processes in the modeling. The Markovian formulation naturally comes into play, and the underlying systems can be modeled via the use of processes of pure jump type. As a result, finite-state Markov chains are used frequently. Very often, the underlying systems have large state spaces with many elements. To obtain the desired optimal control policies or to carry out optimization tasks, the computation required is quite intensive. Therefore, a straightforward implementation of numerical schemes could render the computation infeasible. It is imperative to find systematic approximation procedures to reduce the complexity of the underlying systems. One of the possible resolutions is the use of hierarchical decomposition. In fact, Ando and Fisher proposed the so-called nearly completely decomposable matrix models; see [3] and the references therein. Such a notion has subsequently been applied to queueing networks for organizing the resources in a hierarchical fashion, in computer systems for aggregating memory levels, in economics for reduction of complexity of complex systems (see [3, 24]), and in manufacturing systems for hierarchical production of planning (see [23]). The main idea is based on the following observation: Within a large network, different states often vary at different rates; some of them change rapidly and others vary slowly. To capture such a feature, we introduce a small parameter into the system to highlight the contrasts of different rates of changes. Then the system is naturally decomposed into a number of subsystems with smaller dimensions. We can aggregate the states in each of the subsystem into one state. Using this aggregation scheme, we can further obtain a limit system that is averaged out with respect to certain invariant measures. Consequently, we are able to reduce the complexity of the underlying system substantially.

---

*This research was supported in part by by Wayne State University, by the National Science Foundation under grant DMS-9877090, and by the USAF Grant F30602-99-2-0548 and ONR Grant N00014-96-1-0263, respectively.

To utilize the hierarchical approach for stochastic networks, it is crucial that we have a thorough understanding of the structural properties of the Markov chains. Continuous-time Markovian models were treated in [8, 11, 12, 18, 20, 21, 22, 23, 29, 31, 32, 33, 34], and two-time-scale approaches and singular perturbations for discrete-time systems were used in various applications in telecommunications, control, and optimization problems [1, 2, 7, 17, 25]. Based on our recent results [30] (see also [16] and the references therein for the problems arising from large-dimensional Markov decision problems), we have further studied asymptotic properties of singularly perturbed Markov chains in discrete time. In this paper, we review some of our recent work. We begin with the development of asymptotic expansion of probability vector, proceed with the examination of unscaled and scaled occupation measures, and continue with the limit results of switching diffusion limit. The results obtained will be useful for the subsequent study in control and optimization of large-scale systems and Markov decision processes. We then setup the framework of nearly optimal control of large-scale dynamic systems.

The rest of the paper is arranged as follows. Section 2 presents the formulation of singularly perturbed Markov chains. Section 3 is devoted to various asymptotic results. Section 4 concludes the paper with some further remarks.

2 Singularly Perturbed Markov Chains

Let T be a positive real number. Suppose that $\varepsilon > 0$ is small parameter, and α_k^ε, for $0 \le k \le \lfloor T/\varepsilon \rfloor$, is a discrete-time Markov chain with finite state space $\mathcal{M} = \{1, \ldots, m\}$, where $\lfloor z \rfloor$ denotes the integer part of a real number z. For simplicity, in what follows, we often simply write T/ε and omit the $\lfloor T/\varepsilon \rfloor$ notation. Throughout the paper, for a discrete-time singularly perturbed process, we work with the time horizon $0 \le k \le T/\varepsilon$, whereas for a continuous-time process, we work with the time horizon $[0, T]$. The transition probability matrix of α_k^ε is given by

$$P^\varepsilon(k) = P(k) + \varepsilon Q(k) \tag{2.1}$$

where for each $0 \le k \le T/\varepsilon$, $P(k)$ itself is a transition probability matrix and $Q(k) = (q_{ij}(k))$ is a generator, i.e., for each $i \ne j$, $q_{ij}(k) \ge 0$ and for each i, $\sum_j q_{ij}(k) = 0$ or in short, $Q(k)\mathbb{1}_m = 0$. Hereafter, $\mathbb{1}_\iota \in \mathbb{R}^{\iota \times 1}$ denotes a column vector with all components being 1.

The motivation of such a model stems from various applications in stochastic networks and manufacturing systems [23, 29]. Let us consider a continuous-time Markov chain $\alpha^\varepsilon(t)$ with state space $\mathcal{M} = \{1, \ldots, m\}$ and generator $\widetilde{Q}/\varepsilon + \widehat{Q}$, where $\widetilde{Q} = (\tilde{q}_{ij})$ and

$\widehat{Q} = (\hat{q}_{ij})$ are generators of some Markov chains with stationary transition probabilities. Then $p^\varepsilon(t) = (P(\alpha^\varepsilon(t) = 1), \dots, P(\alpha^\varepsilon(t) = m))$ satisfies the forward equation

$$\dot{p^\varepsilon}(t) = p^\varepsilon(t)\left(\frac{1}{\varepsilon}\widetilde{Q} + \widehat{Q}\right), \quad p(0) = p_0.$$

Introducing a stretched variable $\tau = t/\varepsilon$, $p(\tau) = p^\varepsilon(t)$ satisfies the rescaled forward equation

$$\frac{dp}{d\tau} = p(\tau)(\widetilde{Q} + \varepsilon\widehat{Q}).$$

Let $q_0 = \max_i\{|\tilde{q}_{ii}|\}$. Discretize this differential equation with step size δ such that $1/\delta > q_0$. Then, we have $p((k+1)\delta) = p(k\delta)\widetilde{P}^\varepsilon$, where

$$\widetilde{P}^\varepsilon = (I + \delta(\widetilde{Q} + \varepsilon\widehat{Q})) = (I + \delta\widetilde{Q}) + \varepsilon(\delta\widehat{Q})$$

is a transition probability matrix having the form (2.1). The focus of the current paper is on the structural properties of the discrete-time Markov chains $\widetilde{\alpha}_k^\varepsilon$. The above discussion illustrates the point of view of discretization. Discrete-time singularly perturbed Markov chains also arise in many discrete-time Markov decision processes and control and optimization problems directly.

It is easily seen that in (2.1), $P(k)$ is the dominating force when ε is small, so its structure is important. In accordance with the well-known Kolmogorov classification of states, in a finite-state Markov chain, there is at least one recurrent state, and either all states are recurrent or there is also a collection of transient states in addition to the recurrent states. Furthermore, (see for example, [9, p. 94]) any transition probability matrix of a finite-state Markov chain can be put into the form of either

$$P(k) = \operatorname{diag}(P_1(k), \dots, P_l(k)), \tag{2.2}$$

or

$$P(k) = \begin{pmatrix} P_1(k) & & & & \\ & P_2(k) & & & \\ & & \ddots & & \\ & & & P_l(k) & \\ P_{*,1}(k) & P_{*,2}(k) & \cdots & P_{*,l}(k) & P_*(k) \end{pmatrix}, \tag{2.3}$$

where for each $i \le l$, $P_i(k)$ is a transition matrix within the ith recurrent class, and the last row $(P_{*,1}(k), \cdots P_{*,l}(k), P_*(k))$ in (2.3) corresponds to the transient states. [Here and hereafter, $\operatorname{diag}(A_1, \dots, A_j)$ denotes a block diagonal matrix with matrix entries $A_1, \dots, A_j$ having appropriate dimensions.] A Markov chain with transition matrix given by (2.2) consists of l recurrent classes, whereas a Markov chain with transition matrix (2.3) has l recurrent classes and a number of transient states. For simplicity, the rest of the paper is devoted to Markov chains with stationary transition probabilities, i.e., α_k^ε has transition matrix given by (2.1) with $P(k) = P$, although the full power of our approach is more pronounced for nonstationary models.

To give further motivation, consider a discrete-time linear regulator problem. The traditional setting usually confines to a model with fixed coefficients. To take the robustness into consideration, we introduce the Markovian formulation such that the system matrices $A(\cdot)$, $B(\cdot)$ etc. are all allowed to take values in a finite set. To be more specific, let α_k be a discrete-time Markov chain having finite state space $\mathcal{M}$ ($\mathcal{M}$ contains m elements). Our objective is to find an optimal control u to minimize an expected quadratic cost function. It is well known that the optimal control may be obtained by using the dynamic programming approach. In fact, we can obtain m associated Riccati equations satisfied by the value function. Nevertheless, various stochastic networks have a large state space, i.e., the Markov chain involved has a large number of states. When m is large the amount of computation becomes very intensive and infeasible. Although the system involves a large number of states, different states vary at very different rates. To take advantage of this fact, we deliberately introduce a small parameter $\varepsilon > 0$ into the system. Write α_k^ε in lieu of α_k. Assume that the transition probability matrix has the form (2.1) in which the matrix P takes the form of (2.2) or (2.3). Now the formulation becomes: Suppose that a dynamic system is given by

$$\begin{aligned} x_{k+1} &= x_k + \varepsilon A(\alpha_k^\varepsilon)x_k + \varepsilon B(\alpha_k^\varepsilon)u_k + \sqrt{\varepsilon}w_k, \\ x_0 &= x, \quad \text{a deterministic vector,} \end{aligned} \tag{2.4}$$

where $x_k \in \mathbb{R}^{n_1}$ is the state, $u_k \in \mathbb{R}^{n_2}$ is the control, $A(i) \in \mathbb{R}^{n_1 \times n_1}$ and $B(i) \in \mathbb{R}^{n_1 \times n_2}$ are well defined and have finite values for each $i \in \mathcal{M}$, and $\{w_k\}$ is a sequence of random variables with zero mean. Suppose that $M(i)$, $N(i)$, and D are symmetric such that $M(i)$, D are nonnegative definite and $N(i)$ is positive definite. Define a cost function by

$$J_n^\varepsilon(x, \alpha, u(\cdot)) = E\Big\{\varepsilon \sum_{k=0}^{n-1} [x_k' M(\alpha_k^\varepsilon)x_k + u_k' N(\alpha_k^\varepsilon)u_k] + x_n' D x_n\Big\}, \tag{2.5}$$

where $E = E_{x,\alpha}$ is the expectation given $\alpha_0^\varepsilon = \alpha$ and $x_0 = x$. We aim at finding an optimal control u to minimize the expected quadratic cost function $J_{T/\varepsilon}^\varepsilon(x, \alpha, u(\cdot))$. Using the asymptotic properties of the singularly perturbed Markov chain, we can aggregate the states in each recurrent class as one state leading to a reduced or averaged system in which the total number of Riccati equations reduces to l. If $m \gg l$, a substantial reduction of complexity can be achieved.

3 Asymptotic Results

3.1 Asymptotic Expansion

We construct matched asymptotic expansions of the probability vector $p^\varepsilon(k)$ of the form

$$p^\varepsilon(k) = U^\varepsilon(\varepsilon k) + V^\varepsilon(k) + e^\varepsilon(k), \tag{3.1}$$

where $e^\varepsilon(k)$ represents the approximation error, and

$$U^\varepsilon(\varepsilon k) = \sum_{j=0}^{n} \varepsilon^j \varphi^{\{j\}}(\varepsilon k), \quad V^\varepsilon(k) = \sum_{j=0}^{n} \varepsilon^j \psi^{\{j\}}(k). \tag{3.2}$$

The term $U^\varepsilon(\varepsilon k)$ is the outer expansion and $V^\varepsilon(k)$ the boundary-layer correction. We first construct the formal expansions $U^\varepsilon(\varepsilon k)$ and $V^\varepsilon(k)$, and then prove the error estimate $e^\varepsilon(k) = O(\varepsilon^{n+1})$ uniformly in $0 \le k \le T/\varepsilon$. We need the following assumption:

(A) For both P given by (2.2) and (2.3), P^ε and P are transition probability matrices; for each $i \le l$, P_i is irreducible and aperiodic. In addition, for P given by (2.3), P_* is a matrix having all of its eigenvalues inside the unit circle.

We are now in a position to present the asymptotic expansion result. The proof begins with a construction of the asymptotic expansion terms involved followed by a detailed error estimate. For further development, we refer the reader to [30].

Theorem 3.1 *Suppose that condition* (A) *holds for* P *given by* (2.2) *or* (2.3). *Then the asymptotic expansion can be constructed such that for* $i = 0, \ldots, n$, $\varphi^{\{i\}}(\tau)$ *are sufficiently smooth and* $|\psi^{\{i\}}(k)| \le K\lambda^k$ *for some* $0 < \lambda < 1$. *Furthermore,*

$$\sup_{0\le k\le T/\varepsilon} \left| p^\varepsilon(k) - \sum_{j=0}^{n} \varepsilon^j \phi^{\{j\}}(\varepsilon j) - \sum_{j=0}^{n} \varepsilon^j \psi^{\{j\}}(k) \right| = O(\varepsilon^{n+1}).$$

For ease of presentation, we concentrate on the recurrent case (i.e., P specified by (2.2) in what follows. As a consequence of Theorem 3.1, we have the following corollary.

Corollary 3.2. *Assume condition* (A). *Then the following assertions hold:*

(1) *For the probability distribution vector* $p^\varepsilon(k)$, *we have*

$$p^\varepsilon(k) = \theta(\varepsilon k)\mathrm{diag}(\nu^1, \ldots, \nu^l) + O(\varepsilon + \lambda^k) \tag{3.3}$$

for some λ *with* $0 < \lambda < 1$, *where* ν^i *is the stationary distribution corresponding to the transition matrix* P_i, *and* $\theta(\varepsilon k) = (\theta_1(\varepsilon k), \ldots, \theta_l(\varepsilon k)) \in \mathbb{R}^{1\times l}$ *satisfies*

$$\frac{d\theta(\tau)}{d\tau} = \theta(\tau)\overline{Q}, \quad \theta(0) = x_0 \widetilde{1\!\!1}$$

where $\tau = \varepsilon k$, *and*

$$\overline{Q} = \mathrm{diag}(\nu^1, \ldots, \nu^l) Q \widetilde{1\!\!1}, \quad \widetilde{1\!\!1} = \mathrm{diag}(1\!\!1_{m_1}, \ldots, 1\!\!1_{m_l}). \tag{3.4}$$

(2) *For* $k \le T/\varepsilon$, *the* k*-step transition probability matrix* $(P^\varepsilon)^k$ *satisfies*

$$(P^\varepsilon)^k = \Phi(\varepsilon k) + \varepsilon\widehat{\Phi}(\varepsilon k) + \Psi(k) + \varepsilon\widehat{\Psi}(k) + O\left(\varepsilon^2\right), \tag{3.5}$$

where

$$\begin{aligned} &\Phi(\tau) = \widetilde{1\!\!1}\Theta(\tau)\mathrm{diag}(\nu^1, \ldots, \nu^l) \\ &\frac{d\Theta(\tau)}{d\tau} = \Theta(\tau)\overline{Q}, \quad \Theta(0) = I. \end{aligned} \tag{3.6}$$

3.2 Aggregation

Decompose the state space $\mathcal{M}$ into l subspaces according to the structure of P matrix. That is,

$$\begin{aligned} \mathcal{M} &= \{s_{11}, \ldots, s_{1m_1}\} \cup \{s_{21}, \ldots, s_{2m_2}\} \cup \ldots \cup \{s_{l1}, \ldots, s_{lm_l}\} \\ &= \mathcal{M}_1 \cup \mathcal{M}_2 \cup \cdots \cup \mathcal{M}_l, \end{aligned} \tag{3.7}$$

with $m = m_1 + m_2 + \cdots + m_l$. The subspaces $\mathcal{M}_i$ for $i = 1, \ldots, l$ consist of recurrent states belonging to l recurrent classes. In view of the above decomposition define an aggregated process $\overline{\alpha}_k^\varepsilon$ of α_k^ε by

$$\overline{\alpha}_k^\varepsilon = i \text{ if } \alpha_k^\varepsilon \in \mathcal{M}_i \text{ for } i = 1, \ldots, l. \tag{3.8}$$

Define also continuous-time interpolations with the interpolation interval ε, on $[0, T]$ by

$$\alpha^\varepsilon(t) = \alpha_k^\varepsilon, \quad \text{and} \quad \overline{\alpha}^\varepsilon(t) = \overline{\alpha}_k^\varepsilon \quad \text{if } t \in [\varepsilon k, \varepsilon(k+1)). \tag{3.9}$$

Theorem 3.3. *Assume* (A). *Then as* $\varepsilon \to 0$, $\overline{\alpha}^\varepsilon(\cdot)$ *converges weakly in* $D[0, T]$ *to* $\overline{\alpha}(\cdot)$ *a Markov chain generated by* $\overline{Q}$ *given by* (3.4), *where* $D[0, T]$ *is the space of functions that are right continuous and have left limits endowed with the Skorohod topology.*

3.3 Occupation Measures

Many applications require the understanding of the amount of time the Markov chain spends in a given state. This leads to the study of occupation measures. For $k = 0, \ldots, T/\varepsilon$, $i = 1, \ldots, l$, and $j = 1, \ldots, m_i$, define a sequence of occupation measures

$$\begin{aligned} o^\varepsilon_{k,ij} &= \varepsilon \sum_{r=0}^{k-1} \left(I_{\{\alpha^\varepsilon_r = s_{ij}\}} - \nu^i_j I_{\{\overline{\alpha}^\varepsilon_r = i\}} \right) \\ O^\varepsilon_k &= (o^\varepsilon_{k,11}, \ldots, o^\varepsilon_{k,1m_1}, \ldots, o^\varepsilon_{k,l1}, \ldots, o^\varepsilon_{k,lm_l}). \end{aligned} \tag{3.10}$$

Define a continuous-time interpolation:

$$\begin{aligned} &o^\varepsilon_{ij}(t) = o^\varepsilon_{k,ij} \text{ if } t \in [k\varepsilon, (k+1)\varepsilon) \quad \text{and} \\ &O^\varepsilon(t) = (o^\varepsilon_{11}(t), \ldots, o^\varepsilon_{1m_1}(t), \ldots, o^\varepsilon_{l1}(t), \ldots, o^\varepsilon_{lm_l}(t)). \end{aligned} \tag{3.11}$$

With definition (3.10), we can write O^ε_k recursively as a difference equation

$$\begin{aligned} o^\varepsilon_{k+1,ij} &= o^\varepsilon_{k,ij} + \varepsilon w_{ij}(\alpha^\varepsilon_k) \quad \text{or} \\ O^\varepsilon_{k+1} &= O^\varepsilon_k + \varepsilon W(\alpha^\varepsilon_k), \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} &W(\alpha) = (w_{11}(\alpha), \ldots, w_{1m_1}(\alpha), \ldots, w_{l1}(\alpha), \ldots, w_{lm_l}(\alpha)) \in \mathbb{R}^{1\times m}, \text{ with} \\ &w_{ij}(\alpha) = I_{\{\alpha = s_{ij}\}} - \nu^i_j I_{\{\alpha \in \mathcal{M}_i\}}. \end{aligned} \tag{3.13}$$

Detailed estimate reveals that the following mean square estimate holds.

Theorem 3.4. *Assume* (A). *Then for* $i = 1, \ldots, l$, $j = 1, \ldots, m_i$,

$$\sup_{t\in[0,T]} E|o^\varepsilon_{ij}(t)|^2 = O(\varepsilon).$$

For each $i = 1, \ldots, l, j = 1, \ldots, m_i$, and each $0 < k \leq T/\varepsilon$, define the normalized occupation measure:

$$\begin{aligned} n^\varepsilon_{k,ij} &= \sqrt{\varepsilon} \sum_{r=0}^{k-1} w_{ij}(\alpha^\varepsilon_r) = \frac{1}{\sqrt{\varepsilon}} o^\varepsilon_{k,ij} \\ n^\varepsilon_k &= \left(n^\varepsilon_{k,11}, \ldots, n^\varepsilon_{k,1m_1}, \ldots, n^\varepsilon_{k,l1}, \ldots, n^\varepsilon_{k,lm_l} \right). \end{aligned} \tag{3.14}$$

It is readily seen that

$$n^\varepsilon_{k+1} = n^\varepsilon_k + \sqrt{\varepsilon} W(\alpha^\varepsilon_k). \tag{3.15}$$

To proceed, define a continuous-time interpolation:

$$n^\varepsilon(t) = n^\varepsilon_k \text{ if } t \in [\varepsilon k, \varepsilon(k+1)). \tag{3.16}$$

We can show that $n^\varepsilon(\cdot)$ converges weakly to a switching diffusion. The procedure is as follows:

1. Show that $(n^\varepsilon(\cdot), \bar{\alpha}^\varepsilon(\cdot))$ is tight;
2. verify that the limit of a subsequence of $(n^\varepsilon(\cdot), \bar{\alpha}^\varepsilon(\cdot))$ is a solution to a martingale problem that has a unique solution;
3. characterize the solution of the associated martingale problem;

(4) construct a switching diffusion that is also a solution to the martingale problem and therefore the limit of $(n^\varepsilon(\cdot), \bar{\alpha}^\varepsilon(\cdot))$.

Theorem 3.5. *Assume* (A). *Then* $(n^\varepsilon(\cdot), \overline{\alpha}^\varepsilon(\cdot))$ *converges weakly to* $(n(\cdot), \overline{\alpha}(\cdot))$ *such that the limit is solution of the martingale problem with operator* L *given by*

$$Lf(x,j) = \frac{1}{2} \sum_{j_1,j_2=1}^{m_j} a_{j_1 j_2}(s,j) \frac{\partial^2 f(x,j)}{\partial x_{jj_1} \partial x_{jj_2}} + \overline{Q} f(x,\cdot)(j), \tag{3.17}$$

and the covariance of the limit process is $S(j) = \Sigma(j)\Sigma'(j)$, *where*

$$\Sigma(j) = \mathrm{diag}(0_{m_1 \times m_1}, \ldots, 0_{m_{j-1} \times m_{j-1}}, \sigma(j), 0_{m_{j+1} \times m_{j+1}}, \ldots, 0_{m_l \times m_l}), \tag{3.18}$$

$\sigma(j) \in \mathbb{R}^{m_j \times m_j}$ *satisfying* $\sigma(j)\sigma'(j) = A(j) = (a_{j_1 j_2}(j))$, *and*

$$A(j) = \nu^j_{\mathrm{diag}} \Psi(0,j) + \nu^j_{\mathrm{diag}} \sum_{k=1}^{\infty} \Psi(k,j) + \sum_{k=1}^{\infty} \Psi'(k,j) \nu^j_{\mathrm{diag}}, \tag{3.19}$$

with $\nu^j_{\mathrm{diag}} = \mathrm{diag}(\nu^j_1, \ldots, \nu^j_{m_j}) \in \mathbb{R}^{m_j \times m_j}$, *and* $\Psi(k,j) = (I - \mathbb{1}_{m_j} \nu^j)(P_j)^k$.

4 Further Remarks

We have presented several results concerning the asymptotic properties of singularly perturbed Markov chains in discrete time. As indicated before, the presentation is confined to the stationary case. However, our approach is more useful for nonstationary cases.

One of our current effort is on obtaining nearly optimal controls of large-scale systems driven by singularly perturbed Markov chains. To be more specific, consider the following model. Let α^ε_k be a singularly perturbed Markov chain with a finite-state space $\mathcal{M}$ with transition matrix given by (2.1). Consider a stochastic dynamical system with the state $x^\varepsilon_k \in \mathbb{R}^n$, and control $u_k \in \Gamma \subset \mathbb{R}^{n_1}$. Let

$$\begin{aligned} f(\cdot,\cdot,\cdot) &: \mathbb{R}^n \times \mathbb{R}^{n_1} \times \mathcal{M} \mapsto \mathbb{R}^n, \\ G(\cdot,\cdot,\cdot) &: \mathbb{R}^n \times \mathbb{R}^{n_1} \times \mathcal{M} \mapsto \mathbb{R}. \end{aligned}$$

The problem of interest is

$$\begin{aligned} &\text{minimize: } J^\varepsilon(u) = \varepsilon E \sum_{0}^{T/\varepsilon} G(x^\varepsilon_k, u_k, \alpha^\varepsilon_k), \\ &\text{subject to: } x^\varepsilon_{k+1} = x^\varepsilon_k + \varepsilon f(x^\varepsilon_k, u_k, \alpha^\varepsilon_k), \ x^\varepsilon_0 = x. \end{aligned} \tag{4.1}$$

To solve the problem, we use the relaxed control formulation.

We will show that although the problem is in discrete time, corresponding to (4.1), there is an associate limit problem:

$$\left\{ \begin{aligned} &\text{minimize:} \\ &\quad J(m) = E \sum_{i=1}^{l} \sum_{j=1}^{m_i} \int_0^T \int G(x(t), \varrho, s_{ij}) m_t(d\varrho) \nu^i_j(t) I_{\{\bar{\alpha}(t)=i\}} dt, \\ &\text{subject to:} \\ &\quad x(t) = x + \sum_{i=1}^{l} \sum_{j=1}^{m_i} \int_0^t \int f(x(s), \varrho, s_{ij}) m_s(d\varrho) \nu^i_j(s) I_{\{\bar{\alpha}(s)=i\}} ds, \end{aligned} \right.$$

where $m(\cdot)$ is a relaxed control (see [15]). In lieu of dealing with the original complex system, consider its corresponding limit and apply the optimal or nearly optimal controls of the limit problem to that of the original problem. The goal is to show that such a procedure leads to near optimality of the original problems. The weak convergence approach in conjunction with the relaxed control representation allows us to use weaker conditions. Under the relaxed control formulation, the system is linear in the control variable, and the compactness is also easily obtainable.

References

[1] M. Abbad, J. A. Filar, and T. R. Bielecki, Algorithms for singularly perturbed limiting average Markov control problems, *IEEE Trans. Automat. Control* **AC-37** (1992), 1421-1425.

[2] G. Blankenship, Singularly perturbed difference equations in optimal control problems, *IEEE Trans. Automat. Control* **T-AC 26** (1981), 911-917.

[3] P. J. Courtois, *Decomposability: Queueing and Computer System Applications*, Academic Press, New York, 1977.

[4] M. H. A. Davis, *Markov Models and Optimization*, Chapman & Hall, London, 1993.

[5] F. Delebecque and J. Quadrat, Optimal control for Markov chains admitting strong and weak interactions, *Automatica* **17** (1981), 281-296.

[6] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, J. Wiley, New York, 1986.

[7] F. C. Hoppensteadt and W. L. Miranker, Multitime methods for systems of difference equations, *Studies Appl. Math.* **56** (1977), 273-289.

[8] A. Il'in, R. Khasminskii, and G. Yin, G. Asymptotic expansions of solutions of integro-differential equations for transition densities of singularly perturbed switching diffusions: Rapid switchings, *J. Math. Anal. Appl.* **238** (1999), 516-539.

[9] M. Iosifescu, *Finite Markov Processes and Their Applications*, Wiley, Chichester, 1980.

[10] R.Z. Khasminskii, On stochastic processes defined by differential equations with a small parameter, *Theory Probab. Appl.* **11** (1966), 211-228.

[11] R. Z. Khasminskii, G. Yin and Q. Zhang, Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains, *SIAM J. Appl. Math.* **56** (1996), 277-293.

[12] R. Z. Khasminskii, G. Yin and Q. Zhang, Constructing asymptotic series for probability distribution of Markov chains with weak and strong interactions, *Quart. Appl. Math.* **LV** (1997), 177-200.

[13] T. G. Kurtz, *Approximation of Population Processes*, SIAM, Philadelphia, PA, 1981.

[14] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[15] H. J. Kushner, *Weak Convergence Methods and Singularly Perturbed Stochastic Control and Filtering Problems*, Birkhäuser, Boston, 1990.

[16] R.H. Liu, Q. Zhang, and G. Yin, Nearly optimal control of singularly perturbed Markov decision processes in discrete time, preprint, 1999.

[17] D. S. Naidu, *Singular Perturbation Methodology in Control Systems*, Peter Peregrinus Ltd., Stevenage Herts, UK, 1988.

[18] Z. G. Pan and T. Başar, H^∞-control of Markovian jump linear systems and solutions to associated piecewise-deterministic differential games, in *New Trends in Dynamic Games and Applications*, G. J. Olsder (Ed.), 61-94, Birkhäuser, Boston, 1995.

[19] A. A. Pervozvanskii and V. G. Gaitsgori, *Theory of Suboptimal Decisions: Decomposition and Aggregation*, Kluwer, Dordrecht, 1988.

[20] R. G. Phillips and P. V. Kokotovic, A singular perturbation approach to modelling and control of Markov chains, *IEEE Trans. Automat. Control* **26** (1981), 1087-1094.

[21] J.R. Rohlicek and A. Willsky, The reduction of Markov generators: an algorithm exposing the role of transient states, *J. Assoc. Compt. Mach.* **35** (1988), 675-696.

[22] P.J. Schweitzer, Perturbation theory and finite Markov chains, *J. Appl. Probab.* **5** (1968), 401-413.

[23] S. P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser, Boston, 1994.

[24] H. A. Simon and A. Ando, Aggregation of variables in dynamic systems, *Econometrica* **29** (1961), 111-138.

[25] D. N. C. Tse, R. G. Gallager, and J. N. Tsitsiklis, Statistical multiplexing of multiple time-scale Markov streams, *IEEE J. Selected Areas Comm.* **13** (1995), 1028-1038.

[26] W. A. Thompson, Jr., *Point Process Models with Applications to Safety and Reliability*, Chapman and Hall, New York, 1988.

[27] H. Yang, G. Yin, K. Yin, and Q. Zhang, Control of singularly perturbed Markov chains: A numerical study, preprint, 2000.

[28] G. Yin and Q. Zhang, Control of dynamic systems under the influence of singularly perturbed Markov chains, *J. Math. Anal. Appl.* **216** (1997), 343-367.

[29] G. Yin and Q. Zhang, *Continuous-time Markov Chains and Applications: A Singular Perturbation Approach,* Springer-Verlag, New York, 1998.

[30] G. Yin and Q. Zhang, Singularly perturbed discrete-time Markov chains, *SIAM J. Appl. Math.*, **61** (2000).

[31] G. Yin, Q. Zhang, and G. Badowski, Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states, *Ann. Appl. Probab.* **10** (2000), 549-572.

[32] G. Yin, Q. Zhang, and G. Badowski, Singularly perturbed Markov chains: Convergence and aggregation, *J. Mult. Anal.* **72** (2000), 208-229.

[33] G. Yin, Q. Zhang, and G. Badowski, Occupation measures of singularly perturbed Markov chains with absorbing states, *Acta Math. Sinica*, **16** (2000), 161-180.

[34] Q. Zhang and G. Yin, On nearly optimal controls of hybrid LQG problems, *IEEE Trans. Automat. Control*, **44** (1999) 2271-2282.

On Fluid Analysis of Queues with Selective Message Discarding Policies

Parijat Dube *and Eitan Altman

Project MISTRAL, INRIA,
2004, Route des Lucioles, B.P.93,
06902, Sophia Antipolis Cedex, France
Parijat.Dube, Eitan.Altman@sophia.inria.fr

Abstract

In this paper we analyse a fluid queue with an on-off Markov modulated fluid process as the arrival process. We call each on-period of the arrival process a message. The queue incorporates a selective message discarding policy to prevent network congestion. We work within the framework of Poisson Counter Driven Stochastic Differential Equations and obtain the moment generating function and hence the probability distribution function of the stationary workload process. We then comment on the stability of this fluid queue.

1 Introduction

In this paper we analyse a fluid queue having some selective message discarding policy. A message (or a frame) is a group of packets. Often, a loss of a single packet in a message can result in the loss of the whole message. Thus packets of a message which arrive after a packet is lost from the same message or has the potential of getting lost, due to buffer overflow, have no use any more and it is advantageous to discard them to avoid the occurrence of future congestion due to useless packets (packets of a corrupted message) traversing the network. Such ideas to prevent network congestion were presented in [3] and [4].

Queues with such policies arise in switches incorporating selective message discarding policies. We work in the framework of fluid queues. The input source is modeled as an on-off Markov modulated source with constant fluid arrival rate in on-state; the server has a constant capacity. We call a message the total fluid arriving in an on-period. The fluid queue adopts the policy of accepting fluid corresponding to only those messages, at whose commencement epoch, the queue length (alternatively the work-load process) is less than some threshold level. We consider the case of an infinite buffer. By commencement epoch of a message we mean the time instant when the modulating process goes from off to on state.

The paper is organised as follows. In Section 2 we describe our model and state the main results. In Section 3 we analyse our model and prove our main results. We do the stability analysis of our queue in Section 4 and finally conclude in Section 5.

---

*The work of this author was partially financed by the French Embassy in India

2 Model and Main Results

Our fluid source is a Markov modulated on-off source. Let θ denote the state of the Markov chain. The source sends fluid at rate h when in on($\theta = 1$) state and at rate 0 when in off($\theta = 0$ state). The off and on periods are exponentially distributed with rate λ_1 and λ_2 respectively. The sojourn time in the on-state corresponds to the length of a message. Thus our message lengths are exponentially distributed with parameter λ_2 and silence time (when there is no fluid coming) between messages is exponentially distributed with parameter λ_1.

The discarding policy is such that if at the commencement epoch of a message the workload process $v(t)$ is less than some threshold, the message is admitted, otherwise not.

The fluid source is input to an infinite buffer served with a rate c. To remove trivialities we assume that $c < h$. Here we admit a message fully if at its starting epoch, the workload v is less than B. We write the dynamics of the system in terms of *Poisson Counter Driven Stochastic Differential Equations* (PCDSDE)[1]. Let N_1 and N_2 be Poisson counters with parameters λ_1 and λ_2. We define a new variable x as the *virtual arrival-process* to the buffer. The process $x(t)$ captures the behavior of the discarding policy. The dynamics of $\theta(t)$, $x(t)$ and $v(t)$ are,

$$d\theta(t) = (-\theta(t) + 1)dN_1 - \theta(t)dN_2$$

$$dx(t) = (-x(t) + 1)dN_1\mathbf{I}(v(t) < B) - x(t)dN_2 \tag{1}$$

$$dv(t) = -c\mathbf{I}(v(t) > 0)dt + hx(t)dt \tag{2}$$

Remark 1 *Observe that $\theta(t)$ is a two-state MMFP with rates λ_1 and λ_2. Since the policy does not accept those messages at whose commencement epoch the workload process is greater than or equal to B, hence even if at time t, θ shifts from off to on state but $v(t) \geq B$, the workload process will continue evolving as if $\theta = 0$. Thus, we define a* virtual arrival-process *$x(t)$ which captures this effect. In fact x is the arrival process being admitted to the queue and is not Markovian. A pictorial view of the dynamics of $\theta(t)$, $x(t)$ and $v(t)$ is provided in Figure 1.*

We will now work with the above equations for finding a recursive formula for the average workload moments $E[v^n]$ for all positive integers n. The Laplace-Stieltjes transform, $V(s)(= E[e^{-sv}])$ of v will then be obtained from which we shall obtain the stationary probability density function, $\rho(v)$ of v. We first present our main results.

Proposition 1 *$V(s)$ is given by*

$$V(s) = \frac{k_2(1-p)}{(1+k_1 s)}\left(\frac{\lambda_2 e^{-Bs}}{\lambda_2 + (h-c)s} - 1\right) + 1 - \frac{shk_1 E[x]}{c(1+k_1 s)} \tag{3}$$

where $p = Prob(v < B)$ and

$$k_1 = \left[\frac{(h-c)}{(\lambda_1+\lambda_2)}\left(1 - \frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right] \quad k_2 = \left[1 - \frac{\lambda_2}{\lambda_1+\lambda_2}\left(1 - \frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right] \tag{4}$$

and

$$E[x] = \frac{\lambda_1}{\lambda_1+\lambda_2}\left(p(1-\frac{c}{h}) + \frac{c}{h}\right)$$

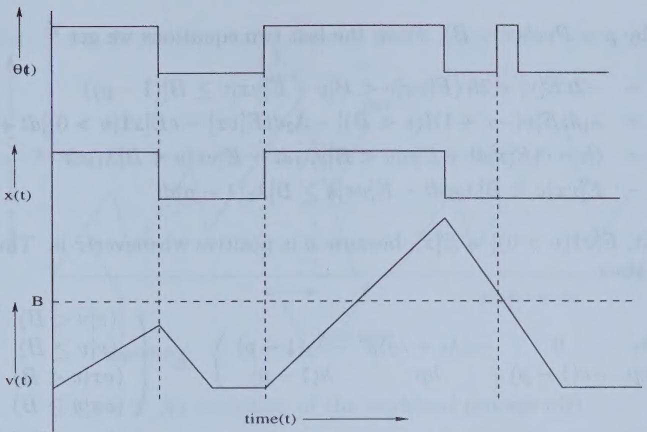


Figure 1: A realization of $\theta(t)$, $x(t)$ and $v(t)$ process

Proposition 2 *The stationary probability density function $\rho(.)$ of $v(t)$is given by,*

$$\rho(m) \;=\; \begin{cases} \rho_1(m) & \text{for } 0 \le m < B \\ \frac{k_2(1-p)\frac{a\lambda_2}{(h-c)}}{\left(a-\frac{\lambda_2}{(h-c)}\right)}\left(e^{-\frac{\lambda_2}{(h-c)}(m-B)} - e^{-a(m-B)}\right) + \rho_1(m) & \text{for } m \ge B \end{cases}$$

where,

$$\rho_1(m) \;=\; \delta(m)\left(1-\frac{h}{c}E[x]\right) + \left(\frac{h}{ck_1}E[x] - \frac{k_2}{k_1}(1-p)\right)e^{-am}$$

k_1, k_2 and $E[x]$ as defined in Proposition 1 with

$$p = \frac{1}{1+\frac{\alpha}{1-\frac{\alpha c}{h}}}$$

where,

$$\alpha \;=\; \frac{\lambda_1 e^{-aB} h}{(\lambda_1+\lambda_2)c}\left(1-k_2+k_2e^{-aB}\right)^{-1}$$

and $a = k_1^{-1}$.

3 Model Analysis and Proofs of the Main Results

Observe that, by using the Itô [1] calculus we get from Equations (1) and (2),

$$\begin{aligned} dv^2(t) &= -2cv(t)\mathbf{I}v(t) > 0)dt + 2v(t)hx(t)dt \\ d(v(t)x(t)) &= v(t)dx(t) + x(t)dv(t) \\ &= v(t)\left((-x(t)+1)dN_1\mathbf{I}(v(t) < B) - x(t)dN_2\right) + x(t)\left(-c\mathbf{I}(v(t) > 0)dt + hxdt\right) \end{aligned}$$

---

[1] For related background refer to the Appendix, see also [1]

We denote by $p = Prob(v < B)$. From the last two equations we get [2]

$$\begin{aligned} dE[v^2] &= -2cE[v] + 2h\left(E[vx|v < B]p + E[vx|v \geq B](1-p)\right) \\ dE[vx] &= \lambda_1 dt E[v(-x+1)\mathbf{I}(v < B)] - \lambda_2 dt E[vx] - cE[x\mathbf{I}(v > 0)]dt + hE[x^2]dt \\ &= (h-c)E[x]dt + E[v|v < B]p\lambda_1 dt - E[vx|v < B]\lambda_1 pdt \\ &- \; E[vx|v < B]\lambda_2 pdt - E[vx|v \geq B]\lambda_2(1-p)dt \end{aligned}$$

Observe that, $E[x\mathbf{I}(v > 0)] = E[x]$, because v is positive whenever x is. Thus if a steady state exists then

$$\begin{pmatrix} p\lambda_1 & 0 & -(\lambda_1+\lambda_2)p & -\lambda_2(1-p) \\ -cp & -c(1-p) & hp & h(1-p) \end{pmatrix} E \begin{pmatrix} (v|v < B) \\ (v|v \geq B) \\ (vx|v < B) \\ (vx|v \geq B) \end{pmatrix} + \begin{pmatrix} (h-c)E[x] \\ 0 \end{pmatrix} = 0 \qquad (5)$$

Thus we shall have two equations in six unknowns (the four conditional expectations, $E[x]$ and p). We shall obtain expressions for $E[v|v > B]$, $E[vx|v > B]$ and $E[x]$ in the following section. Observe also the fact that the fluid level v never remains steady at B (visiting it only at isolated points in time) and hence we take the density function of v to be continuous at B. Note that the density function need not be *differentiable* at B. Thus we guess that the density function will be of the form

$$\rho(m) = \begin{cases} C_0\delta(m) & \text{for } m = 0 \\ \rho_1(m) & \text{for } m \in (0, B) \\ \rho_2(m) & \text{for } m \in [B, \infty) \end{cases}$$

3.1 Calculation of $E[v|v \geq B]$ and $E[vx|v \geq B]$

Observe that when $v(t) \geq B$ and the state of $\theta(t)$ changes from 0 to 1 for the first time (and succeeding times) then the incoming fluid is not accepted. This is done till time t when $v(t) < B$. A realization of the evolution of $v(t)$ is shown in Figure 2. Observe that $v(t)$ can only cross the threshold of B at any time t, if $x(t) = x(t^-) = 1$. Since the sojourn time in a state is exponentially distributed, the excess time the Markov chain spends in state 1 after time t is again exponentially distributed given that the chain was in state 1 at time t. The fluid level will rise steadily with a rate $(h - c)$ from t until the Poisson process N_2 causes x to change from 1 to 0. Then v will start decreasing at a steady rate of c until the buffer level is B. During this period (when $v \geq B$), even if the process θ changes to 1 (and fluid starts arriving at rate h), the fluid is not accepted (the fact captured by the presence of an indicator function $\mathbf{I}(v < B)$ in the dynamics of $x(t)$ by Equation 1).

Let T_i $i = 1, 2, \ldots$ be the random variable denoting the time spent by v after crossing B at the ith cross. Thus

$$T_i = X_i + \frac{X_i(h-c)}{c} = X_i\frac{h}{c} \qquad (6)$$

---

[2] for notational convenience we shall not explicitly show t in parentheses

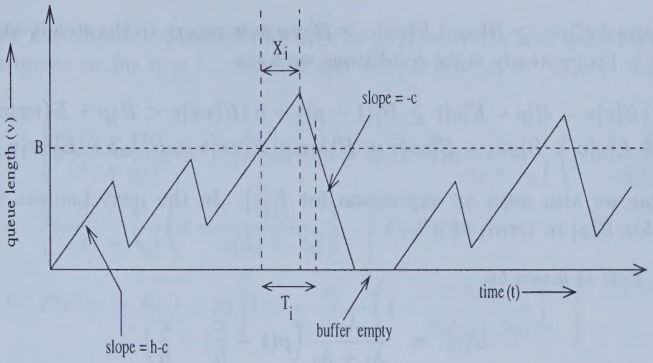


Figure 2: An evolution of the workload process $v(t)$

where X_i is the excess sojourn time in state 1. Notice that the condition $v(t) \geq B$ implies that at the time of crossing the level B, say at time $t_1 \leq t$, $x(t_1) = 1$. Thus, the sequence $\{T_i\}$ is i.i.d. ($X_i \sim \exp \lambda_1$) and the process $T = \sum_{i=1}^{n} T_i$ is a renewal process. We can write

$$\begin{aligned} E[v|v \geq B] &= E[v|v \geq B, \frac{dv}{dt} = (h-c)]P(\frac{dv}{dt} = (h-c)|v \geq B) \\ &+ E[v|v \geq B, \frac{dv}{dt} = -c]P(\frac{dv}{dt} = -c|v \geq B) \end{aligned} \tag{7}$$

Also from Equation (6) we get $X_i = \frac{c}{h}T_i$. Thus we have $P(\frac{dv}{dt} = (h-c)|v \geq B) = \frac{c}{h}$ and $P(\frac{dv}{dt} = -c|v \geq B) = 1 - P(\frac{dv}{dt} = (h-c)|v \geq B) = 1 - \frac{c}{h}$. Also, the process $\{X_i\}$ is i.i.d. and so in the space where $(\frac{dv}{dt} = h-c, v \geq B)$ the process $X = \sum_{i=1}^{n} X_i$ is renewal. By the *Reward Renewal Theorem* [6] we can easily show that,

$$E[v|v \geq B, \frac{dv}{dt} = (h-c)] = B + (h-c)\lambda_2^{-1}$$

And,

$$E[v|v \geq B, \frac{dv}{dt} = -c] = B + (h-c)\lambda_2^{-1}$$

Thus we have from Equation 7

$$E[v|v \geq B] = B + (h-c)\lambda_2^{-1}$$

We will now proceed to calculate $E[vx|v \geq B]$. By similar arguments, we can write

$$E[vx|v \geq B] = E[vx|v \geq B, \frac{dv}{dt} = (h-c)]\frac{c}{h} + E[vx|v \geq B, \frac{dv}{dt} = -c](1 - \frac{c}{h})$$

Observe that, $x = 1$ when $\left(\frac{dv}{dt} = (h-c), v \geq B\right)$ and $x = 0$ when $\left(\frac{dv}{dt} = -c, v \geq B\right)$. Thus,

$$\begin{aligned} E[vx|v \geq B] &= E[v|v \geq B, \frac{dv}{dt} = (h-c)]\frac{c}{h} \\ &= \left(B + (h-c)\lambda_2^{-1}\right)\frac{c}{h} \end{aligned}$$

Having obtained $E[v|v \geq B]$ and $E[vx|v \geq B]$ we now return to the steady-state condition (Equation 5). From steady state conditions we have

$$-c\left(E[v|v < B]p + E[v|v \geq B](1-p)\right) + h\left(E(vx|v < B)p + E(vx|v \geq B)(1-p)\right) = 0$$
$$(h-c)Ex + E[v|v < B]p\lambda_1 - E[vx|v < B]\lambda_1 p - E[vx|v < B]\lambda_2 p - E[vx|v \geq B]\lambda_2(1-p) = 0$$

Observe that we also need an expression for $E[x]$. In the next Lemma we obtain an expression for $E[x]$ in terms of p.

Lemma 1 *$E[x]$ is given by,*

$$E[x] \;=\; \frac{\lambda_1}{\lambda_1+\lambda_2}\left(p(1-\frac{c}{h}) + \frac{c}{h}\right) \tag{8}$$

Proof:
In steady state from Equation 1, we get

$$\lambda_2 E[x] \;=\; p\lambda_1 - E[x|v < B]p\lambda_1 \tag{9}$$

But,

$$\begin{aligned} E[x] &= E[x|v < B]p + E[x|v \geq B](1-p) \\ &= E[x|v < B]p + \Bigg(E[x|v \geq B, \frac{dv}{dt} = h-c]P(\frac{dv}{dt} = h-c|v \geq B) \\ &\quad + E[x|v \geq B, \frac{dv}{dt} = -c]P(\frac{dv}{dt} = -c|v \geq B)\Bigg)(1-p) \\ &= E[x|v < B]p + \frac{c}{h}(1-p) \end{aligned} \tag{10}$$

The last equality follows as $x = 0$ for the case $\left(v \geq B, \frac{dv}{dt} = -c\right)$ and $P(\frac{dv}{dt} = h - c|v \geq B) = \frac{c}{h}$ from preceding discussion. From Equations (9) and (10) we get (8).

3.2 Proof of Proposition 1:

From the steady-state equations, after some calculations we get,

$$\begin{aligned} E[v] &= \left[\frac{h(h-c)}{c(\lambda_1+\lambda_2)}\left(1 - \frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right] E[x] \\ &\quad + E[v|v \geq B](1-p)\left[1 - \frac{\lambda_2}{\lambda_1+\lambda_2}\left(1 - \frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right] \end{aligned} \tag{11}$$

Similarly we can write the steady state equations for any n, $n \geq 1$ by Itô rule. Thus, in general for the nth moment we get [3]

$$E[v^n|v < B] \;=\; \left[n\frac{h(h-c)}{c(\lambda_1+\lambda_2)}E[v^{n-1}x] - E[v^n|v \geq B](1-p)\frac{\lambda_2}{\lambda_1+\lambda_2}\right]\left(p - \frac{hp\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}$$

---

[3] For details see [2]

Observe that we can obtain expressions for $E[v^n|v \geq B]$ and $E[v^n x|v \geq B]$ for any $n > 1$ by similar arguments as for $n = 1$. Also for any $n \geq 1$, $E[v^n x] = \frac{c}{h}E[v^n]$. Thus we get for $n \geq 2$

$$\begin{aligned}
E[v^n|v < B] &= \left[n\frac{(h-c)}{(\lambda_1+\lambda_2)}E[v^{n-1}] - E[v^n|v \geq B](1-p)\frac{\lambda_2}{\lambda_1+\lambda_2}\right]\left(p - \frac{hp\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1} \\
\Rightarrow E[v^n] &= \left[n\frac{(h-c)}{(\lambda_1+\lambda_2)}\left(1-\frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right]E[v^{n-1}] \\
&+ \quad E[v^n|v \geq B](1-p)\left[1 - \frac{\lambda_2}{\lambda_1+\lambda_2}\left(1-\frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right] \qquad (12)
\end{aligned}$$

and for $n = 1$

$$\begin{aligned}
E[v^n] &= \left[n\frac{h(h-c)}{c(\lambda_1+\lambda_2)}\left(1-\frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right]E[v^{n-1}x] \\
&+ \quad E[v^n|v \geq B](1-p)\left[1 - \frac{\lambda_2}{\lambda_1+\lambda_2}\left(1-\frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right] \qquad (13)
\end{aligned}$$

Thus we have obtained recursive relation between $E[v^n]$ and $E[v^{n-1}]$ for $n \geq 1$ in Equations (12) and (13). Observe that, for $n \geq 0$, $E[v^n|v \geq B] = E[v^n|v \geq B, \frac{dv}{dt} = (h-c)] = E[v^n|v \geq B, \frac{dv}{dt} = -c]$, hence we calculate

$$E[v^n|v \geq B] \;=\; E\left[\frac{(B + X_1(h-c))^{n+1} - B^{n+1}}{(n+1)(h-c)}\right]\lambda_2 \qquad (14)$$

Substituting the expression for $E[v^n|v \geq B]$ from Equation (14) in Equation (12) and then multiplying both the sides of the resulting equation by $\frac{(-s)^n}{n!}$ and summing from $n = 2$ to ∞, we can write,

$$\begin{aligned}
\sum_{n\geq 2}\frac{(-s)^n}{n!}E[v^n] &= \left[\frac{(h-c)}{(\lambda_1+\lambda_2)}\left(1-\frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right](-s)\sum_{n\geq 2}\frac{(-s)^{n-1}}{n-1!}E[v^{n-1}] \\
&+ \quad (1-p)\left[1 - \frac{\lambda_2}{\lambda_1+\lambda_2}\left(1-\frac{h\lambda_1}{c(\lambda_1+\lambda_2)}\right)^{-1}\right]E\left[\frac{1}{(h-c)(-s)}\sum_{n\geq 2}\right. \\
&\left.[(B + X_1(h-c))^{n+1} - B^{n+1}]\frac{(-s)^{n+1}}{n+1!}\right]\lambda_2 \qquad (15)
\end{aligned}$$

Adding to Equation (15) $(-s) \times$ *Equation* (13) we get,

$$\begin{aligned}
\sum_{n\geq 1}\frac{(-s)^n}{n!}E[v^n] &= k_1(-s)\sum_{n\geq 2}\frac{(-s)^{n-1}}{n-1!}E[v^{n-1}] + (1-p)k_2 E\left[\frac{1}{(h-c)(-s)}\sum_{n\geq 1}\right. \\
&\left.[(B + X_1(h-c))^{n+1} - B^{n+1}]\frac{(-s)^{n+1}}{(n+1)!}\right]\lambda_2 + \frac{hk_1}{c}E[x](-s) \qquad (16)
\end{aligned}$$

where k_i are given in Equation (4). First we will look at the second expectation term on the right side of Equation (16).

$$E\left[\frac{1}{(h-c)(-s)}\sum_{n\geq 1}[(B + X_1(h-c))^{n+1} - B^{n+1}]\frac{(-s)^{n+1}}{n+1!}\right] = E\left[\frac{e^{-Bs}}{(h-c)s}\left(1 - e^{-X_1(h-c)s}\right) - X_1\right]$$

From which we get,

$$
\begin{aligned}
E\left[\frac{e^{-Bs}}{(h-c)s}\left(1-e^{-X_1(h-c)s}\right)-X_1\right] &= \frac{e^{-Bs}}{(h-c)s}\lambda_2\int_0^\infty\left(1-e^{-x(h-c)s}\right)e^{-\lambda_2 x}dx-\lambda_2^{-1} \\
&= \frac{e^{-Bs}}{\lambda_2+(h-c)s}-\frac{1}{\lambda_2}
\end{aligned}
$$

Using $V(s)=\sum_{n\geq 0}\frac{(-s)^n Ev^n}{n!}$ we get Equation (3) from Equation (16).

3.3 Proof of Proposition 2:

From Proposition (1) we have

$$
V(s)=\frac{k_2(1-p)\lambda_2}{k_1(h-c)}e^{-Bs}\frac{1}{(\frac{\lambda_2}{(h-c)}+s)(s+k_1^{-1})}-\frac{k_2(1-p)}{k_1(s+k_1^{-1})}+1-\frac{h}{c}E[x]+\frac{hE[x]}{ck_1(k_1^{-1}+s)} \tag{17}
$$

Also viewing the first term on the right side of Equation (17) as the product of two LSTs $\frac{1}{s+\frac{\lambda_2}{(h-c)}}$ and $\frac{1}{s+k_1^{-1}}$ and denoting by L^{-1} the inverse LST [4] we get (with $*$ denoting *convolution* and $a=k_1^{-1}$)

$$
\begin{aligned}
L^{-1}\left(\frac{1}{\left(s+\frac{\lambda_2}{(h-c)}\right)(s+a)}\right) &= e^{-\frac{\lambda_2}{(h-c)}m} * e^{-am} \\
&= \frac{1}{\left(a-\frac{\lambda_2}{(h-c)}\right)}\left(e^{-\frac{\lambda_2}{(h-c)}m}-e^{-am}\right)
\end{aligned}
$$

Thus for $m<B$ we have, (from the inverse of *Laplace-Stieltjes Transform* Equation (17)),

$$
\rho(m)=\rho_1(m) \quad = \quad \delta(m)\left(1-\frac{hE[x]}{c}\right)+\left(\frac{h}{ck_1}E[x]-\frac{k_2}{k_1}(1-p)\right)e^{-am} \tag{18}
$$

And for $m>B$

$$
\rho(m)=\rho_2(m) \quad = \quad \frac{k_2(1-p)a\frac{\lambda_2}{(h-c)}}{\left(a-\frac{\lambda_2}{(h-c)}\right)}\left(e^{-\frac{\lambda_2}{(h-c)}(m-B)}-e^{-a(m-B)}\right)+\rho_1(m) \tag{19}
$$

where, $\rho_2(m)$ is the continuous density function for $m\in[B,\infty)$. Observe that Equation (19) involves a term $1-p$ on the right side where $1-p=P(m\geq B)$. However $1-p=\int_B^\infty\rho_2(m)dm$. Thus we can write for $B>0$,

$$
\begin{aligned}
1-p &= \int_B^\infty\left[\frac{k_2(1-p)a\frac{\lambda_2}{(h-c)}}{\left(a-\frac{\lambda_2}{(h-c)}\right)}\left(e^{-\frac{\lambda_2}{(h-c)}(m-B)}-e^{-a(m-B)}\right)+\left(\frac{h}{ck_1}E[x]-ak_2(1-p)\right)e^{-am}\right]dm \\
&= \frac{k_2(1-p)a\frac{\lambda_2}{(h-c)}}{\left(a-\frac{\lambda_2}{(h-c)}\right)}\left(\frac{1}{\frac{\lambda_2}{(h-c)}}-\frac{1}{a}\right)+\left(\frac{h}{ck_1}E[x]-ak_2(1-p)\right)\frac{e^{-aB}}{a} \\
&= k_2(1-p)+(\frac{h}{c}E[x]-k_2(1-p))e^{-aB}
\end{aligned}
$$

---

[4] For a random variable X with distribution $F(x)$ and LST $L(s)(=\int_0^\infty e^{-sx}dF(x)dx)$ by L^{-1} we mean,

$$
L^{-1}[L(s)]=\frac{dF(x)}{dx}
$$

Thus we get solving for $(1-p)$

$$1-p \;=\; \frac{e^{-aB}hE[x]}{c}\left(1-k_2+k_2e^{-aB}\right)^{-1} \tag{20}$$

However we have the expression for $E[x]$ from Equation (8) in terms of p. Thus from Equation (8) and (20), solving for p gives,

$$p \;=\; \frac{1}{1+\frac{\alpha}{1-\frac{\alpha c}{h}}}$$

where, $\alpha = \frac{\lambda_1 e^{-aB}h}{(\lambda_1+\lambda_2)c}\left(1-k_2+k_2e^{-aB}\right)^{-1}$. Having obtained p we have the density function $\rho_1(.)$ and $\rho_2(.)$.

4 Stability Analysis

Next, we show that the workload process is a renewal process. To that end we consider the Markov chain V_n which is the workload process at the nth transition of $\theta(t)$ from state 0 to state 1. In other words, it is the workload as seen by the commencement of the nth potential arriving message. We shall show that V_n is a Harris recurrent Markov chain, and that the empty state is recurrent whenever $c > 0$. This will then imply that the original workload process is a renewal process.

To that end, we recall the following sufficient condition for the Harris recurrence which follows from Theorem 14.0.1 in p. 330 of [5].

Lemma 2 *Assume that V_n is Ψ-irreducible (for some Ψ) and aperiodic. Let there be a function f, some $\epsilon > 0$, and some small set C such that*

$$E[f(V_{n+1}) - f(V_n)|V_n = v] < -\epsilon + \mathbf{I}\{v \notin C\}. \tag{21}$$

Then the Markov chain V_n is positive recurrent, it has a stationary probability π, and the n-step transition probabilities P^n converges to the π in total variation as $n \to \infty$.

In the above Lemma, a set C is *small* if there exists some integer n, a constant $g > 0$ and a probability measure ϕ over the state space R such that

$$[P^n]_{xA} > g\phi(A) \tag{22}$$

for all $x \in C$ and measurable set $A \subset R$.

Theorem 1 *Assume that $c > 0$. Then the process V_n is Harris recurrent.*

Proof: Define $C = [0, B]$ and $f(v) = v$. Let T be an exponentially distributed random variable with parameter λ_2. Then for all $v \notin C$,

$$E[V_{n+1} - V_n|V_n = v] < -cE[\min(T, B/c)] =: -\epsilon.$$

Thus (21) holds. Next we check that C is indeed a small set. Let Y_i be an exponentially random variable with parameter λ_i, $i = 1, 2$. Viewing Y_1 as the length of the off-period and Y_2 as the 0 of the on-period, we have for any $v \in C$ and measurable $A \subset R$

$$[P^2]_{vA} \geq P(\frac{B}{h-c} < Y_2 < \frac{2B}{h-c}, Y_1 > \frac{3B}{c})P_{oA}.$$

Thus, (22) holds with $\phi(A) = P_{0A}$ and $g = P(\frac{B}{h-c} < Y_2 < \frac{2B}{h-c}, Y_1 > \frac{3B}{c})$.

Observe that V_n is ϕ-irreducible (p. 70,87 of [5]) since the probability of eventually reaching any measurable set $A \subset R$ from any state x is greater that $g\phi(A)$. Finally, the aperiodicity follows since the probability to go from state 0 to state 0 is strictly positive.

5 Conclusion

In this paper we analysed an infinite buffer fluid queue which employs a threshold policy for message discarding. We obtain the exact Laplace-Stieltjes transform, $V(s)$ and hence the exact stationary probability density function, $\rho(.)$ of the stationary workload process. We then analysed the stability of this fluid queue and established that the Markov chain V_n(workload process at the nth transition of the source state from off to on) is Harris recurrent. We are currently working on finding exact expressions for the performance measures like goodput, throughput etc.

Appendix

Consider a stochastic differential equation driven by n independent Poisson Counters $N_1, \ldots, N_n$,

$$dx = f(x)dt + \sum_{i=1}^{n} g_i(x)dN_i, \quad x \in R^n$$

Then we have the following Itô rule:
If $\phi : R^n \to R$ is differentiable function, then

$$d\phi(t) = < \frac{\partial \phi}{\partial x}, f(x) > dt + \sum_{i=1}^{n}[\phi(x(t) + g_i(x(t))) - \phi(x(t))]dN_i$$

Also, since $x(t)$ is continuous from the left and the Poisson counter is taken to be continuous from the right, we have

$$\frac{d}{dt}(E[x(t)]) = E[f(x(t))] + \sum_{i=1}^{n}(E[g_i(x(t)), t])\lambda_i$$

References

[1] Wei-Bo Gong Roger W. Brockett and Yang Guo. New analytical methods for queueuing systems. *Proc. of IEEE Conference on Decision and Control, Phoenix, Arizona, USA*, 1999.

[2] Parijat Dube and Eitan Altman. On Fluid Analysis of Queues with Selective Message Discarding Policies. *INRIA Research Report (available on request)*, 2000.

[3] S. Floyd and V. Jacobson. Random Early Detection Gateways for Conjestion Avoidance. *IEEE/ACM Transactions on Networking*, 1(4):25–39, 1993.

[4] S. Floyd and A. Romanow. Dynamics of TCP traffic over ATM networks. *Proc. of ACM SIGCOMM'94*, pages 79–88, September 1994.

[5] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, 1993.

[6] Ronald Wolff. *Stochastic Modeling and Theory of Queues.* Prentice Hall, NJ, 1989.

Equivalent Models and Analysis for Multi-Stage Tree Networks of Deterministic Service Time Queues

Michael J. Neely
MIT--LIDS
mjneely@mit.edu

Charles E. Rohrs
Tellabs Research Center and MIT
crohrs@trc.tellabs.com

Abstract -- **Tree networks of single server, deterministic service time queues are often used as models for packet flow in data communication systems with Asynchronous Transfer Mode (ATM) traffic. In this paper, we present a method for analyzing packet occupancy in these systems without making any assumptions about the nature of the underlying input processes. We demonstrate how analysis of these multi-stage tree systems can be reduced to the analysis of a much simpler 2-stage equivalent model. We also develop an expression for first moments of queue occupancy in terms of first moments of a simple 1-stage equivalent model. From this, we observe an interesting phenomenon for general types of *distributable inputs*: Expected occupancy at any interior queue within a multi-stage tree network is a concave function of the multiple exogenous input rates. Expected occupancies in nodes on the edge of the network are shown to be convex.**

I. INTRODUCTION

In this paper we analyze the packet occupancy in multi-input, single output tree systems of deterministic service time queues (Fig. 1). Such systems are often used as models for packet flow in networks with ATM traffic, where packets are broken into fixed bit-length *cells*. These cells have a service time proportional to their bit-length. Assuming that the processing speed at each node is the same, the cells thus have a deterministic service time T in each node. In this paper, we shall continue to use the term "packets" rather than "cells," keeping in mind that, unless otherwise stated, the packets are fixed in length and have deterministic service time T.

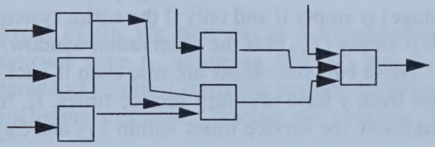


Figure 1: A multi-stage tree network of queues.

Much of the previous work in this area has concentrated on tandem chains of queues and memoryless inputs ([1]-[4]). In [5], an average queue length analysis was performed for a more general tree network using the properties of memoryless inputs. Their approach relied on a lemma from [6] that used combinatorial analysis. The approach here is applicable to more general input processes and the analysis here is self-contained and applicable to more than the mean queue length. We also work in continuous time, which enables exact analysis even when arrival streams are asynchronous.

We first examine a simple case of such a system: a 2-stage, 2-queue system with arbitrary exogenous packet arrival processes at both the first and second stages. We develop a simpler, one-queue equivalent model of this system and prove that the input-output behavior of the original system is preserved in this equivalent model. This fact was presented previously in [1] and used to analyze waiting times in tandem chains of queues with memoryless inputs. The approach used in this paper was developed independently and differs from that used in [1] as the proof presented here emphasizes the equivalence on every sample path.

We extend the use of this idea to analyze multi-stage tree systems with arbitrary input processes. We decompose the large system into "atomic" blocks of 2-stage, 2-node subsystems, and then use our equivalent model iteratively on each sub-system. The result is a simplified 2-stage system in tandem with a series of time delays or *observation windows*.

If the various input processes are all independent and stationary, the time delays can be ignored, and exact analysis of packet occupancy in a multi-stage network can be reduced to the analysis of a 2-stage equivalent model.

These models considerably reduce the complexity of the original network. They can be analyzed quite generally, and often occupancy characteristics such as means, variances, and aggregate distribution functions can be found (see [7, 8]).

Here, we use this equivalent model to develop an expression for mean occupancy at any node in a multi-stage tree system in terms of the mean occupancy of a single-stage (one queue) system whose inputs are a superposition of the original exogenous inputs. As an example, we develop explicit expressions for the average occupancy in any node when the exogenous inputs are (i) memoryless, and (ii) periodic with independent phases.

Finally, we explore a concavity phenomenon of expected occupancy. For a general class of *distributable inputs*, we show that the average queue length of any interior node in a multi-stage, deterministic service time tree system is a concave function of the exogenous input rates. Expected occupancies in nodes at the edge of the network are convex. This provides insight into buffering requirements and rate assignments in ATM tree networks. Deterministic service time queues at earlier stages smooth traffic for downstream nodes. Unbalanced loadings give the previous stage queues the best situation for smoothing the input process for the next stage.

II. THE 2-STAGE, 2-NODE SYSTEM

Here we analyze the simplest non-trivial example of a multi-stage tree system: the 2-stage, 2-node system of Fig. 2a. Exogenous input streams with arbitrary arrival patterns enter the system both at node 1 and at node 2.

In the sequel, we show that if service times of all packets in nodes 1 and 2 are deterministic with length T, then the system can be modeled as a much simpler 1-queue system in tandem with a pure delay or *observation window.* Under identical sets of inputs, the original system and its equivalent model simultaneously produce the same outputs for all time. This equivalence relationship was discovered independently and previously in [1]. Here we state and prove the theorem with greater detail and generality using sample path arguments. This sample path theory provides a means for understanding and analyzing the larger networks addressed in Section III and beyond.

Consider the system in Fig. 2a where inputs x are sent through an initial queuing stage with deterministic service time T. The departures from this first stage are then added to the external inputs y, and the superposition is sent through the queuing system S_{A2}. The nature of the x, y input arrivals is arbitrary.

Consider the same sequences of packets, x and y, as inputs to System B as shown in Fig. 2b. The queueing nodes S_{A2} and S_{B2} are identical. However, the first stage node that x packets pass through in System A is now replaced in System B by an observation window.

Description of Observation Window: Note that the "observation window" is a passive device that simply observes the number of customers which travel through it during a time interval T. Inputs x enter this window, pass through it, and exit after exactly T seconds without having their relative positions distorted in any way. Hence, the input to S_{B2} coming from the window is simply a time-delayed version of x (Fig. 3). Such a device can also be viewed as a $*/D/\infty$ queue.

The theorem below describes the equivalence of Systems A and B, assuming that both start out completely empty at time t=0.

Theorem 1 (Equivalence Theorem): If the service times within S_{A2} and S_{B2} of customers entering from the x input line are greater than or equal to T, then:

i) The final stage systems S_{A2} and S_{B2} have identical busy and idle periods.

ii) The entire system A (S_{A2} plus the first stage) is empty if and only if the entire system B is empty (S_{B2} plus the observation window).

These two statements are true even if packets from y have arbitrary service times. If, in addition, the service times within S_{A2} and S_{B2}

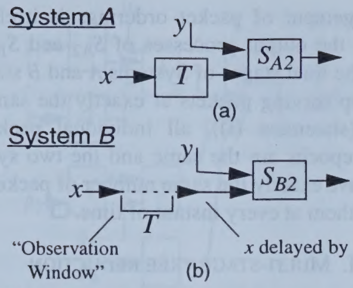


Figure 2: A 2-stage system and its equivalent model. Sub-system S_{A2} is identical to S_{B2}.

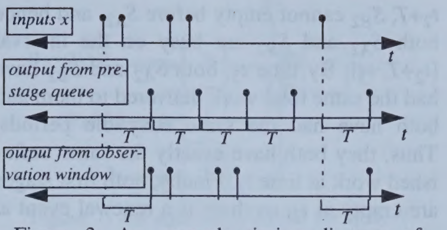


Figure 3: An example timing diagram of inputs x and the corresponding outputs in both the System A pre-stage queue and the System B observation window.

of packets from both input lines x and y are deterministically $\tilde{T}$, where $\tilde{T} \geq T$, then:

iii) The accumulated number of departures from System A and System B is the same at every instant of time, and hence the entire 2-stage System A always has exactly the same number of customers within it as the entire 2-stage System B.

Caveat: Note here that the actual ordering of the packets served may be shuffled in the equivalent model--it is only the departure time epochs that must remain the same for both systems.

Proof of Theorem 1: We start out at time t_1, and suppose that at this time:

1. The first stages of both systems are empty.
2. Up to this point, all time epochs marking the busy and idle periods of the second stage systems S_{A2} and S_{B2} have been identical,
3. The amount of unfinished work left in the final stages of both systems is identical.

We call the event that these three conditions are satisfied simultaneously a *simultaneous renewal event.* Such an event occurs at time 0 when there have been no outputs and both systems are empty. Whether or not another renewal occurs is irrelevant: we simply prove our theorem over the entire (possibly infinite) duration between simultaneous renewal events.

Proof of (i): We start out at a simultaneous renewal time t_1, assuming that S_{A2} and S_{B2} have had identical busy/idle periods up to and including this time. Tracing the timeline of events shown in Fig. 4, we show the busy/idle periods remain synchronized until the next renewal. The next event in the timeline occurs at time t_2 when a busy period of the first stage of System A begins. This busy period consists of n packets $\{x_1, x_2, \ldots, x_n\}$ spaced closely enough together so that the first stage queue in System A does not empty until time t_3 when the n^{th} packet x_n exits the first stage and moves into system S_{A2} (clearly, $t_3 = t_2 + nT$). The packet x_1 that initiates this busy period travels through the first stages of Systems A and B in exactly the same manner, and enters the second stages of both systems at time t_2+T. Since the x, y inputs to systems S_{A2} and S_{B2} are identical during the time interval $[t_1, t_2+T]$, we know that throughout this interval the amount of unfinished work in both S_{A2} and S_{B2} is the same as are the busy and idle periods of S_{A2} and S_{B2}.

Since the service time of x packets in system S_{A2} is at least as long as the service time T in the front end system that feeds into it, system S_{A2} cannot empty before time t_3.

Notice now that because there is no queueing in the observation window, the work delivered to S_{B2} during the interval $[t_2 + T, \tau]$ is greater than or equal to the work delivered to S_{A2} for all $\tau \in [t_2 + T, t_3]$. Thus, after time

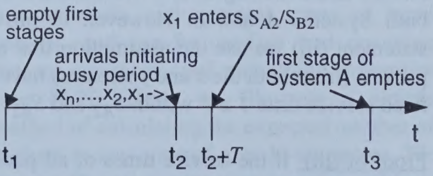

empty first stages
x_1 enters S_{A2}/S_{B2}
arrivals initiating busy period $x_n,\ldots,x_2,x_1$->
first stage of System A empties
t
t_1
t_2
t_2+T
t_3

Figure 4: Events between renewal times t_1 and t

t_2+T, S_{B2} cannot empty before S_{A2}, and hence both S_{A2} and S_{B2} are busy on the interval $[t_2+T, t_3]$. By time t_3, both S_{A2} and S_{B2} have had the same total work delivered to them and both have had the same busy/idle periods. Thus, they both have exactly the same unfinished work at time t_3. Finally, both first stages are empty at t_3, so there is a renewal event at t_3, and (i) is proved. ❑

Proof of (ii): Suppose System B is empty. Thus, S_{B2} is empty, and hence S_{A2} must be empty by statement (i). Since the observation window of System B is empty, no packets have entered from the x input line for at least T seconds. Therefore, no packets have entered the first stage of System A for at least T seconds. Thus, if this first stage now contains packets, then these packets must have been in the system T seconds earlier. But if this is the case, then the packet that was in service at the first stage of System A at that time would have been sent to the second stage S_{A2}, and it would now be in the system S_{A2}. (Recall that service times of x packets in S_{A2} and S_{B2} are no shorter than T seconds). This contradicts the fact that S_{A2} is empty. Hence, the first and second stages of System A are empty.

The converse implication follows from a similar argument, and (ii) is proved. ❑

Notice that up to this point, we have only assumed that packets from the x input stream have a deterministic service time T within the first queuing stage of System A, and that their service times within the second stage systems S_{A2} and S_{B2} are at least as large as T. Thus, statements (i) and (ii) hold even if service times of x packets within S_{A2} and S_{B2} are variable with a minimum time of T seconds. Furthermore, the packets from the y input stream can have any type of arbitrarily distributed service times as long as they are the same for both Systems A and B. However, to prove statement (iii) we use the assumption that all packets (from both the x and y streams) have a fixed service time $\tilde{T} \geq T$ within S_{A2} and S_{B2}.

Proof of (iii): If the service times of all packets in S_{A2} and S_{B2} are identically $\tilde{T}$ seconds, then all packets are homogeneous and any rearrangement of packet orderings is irrelevant to the output processes of S_{A2} and S_{B2}. Since the final stages of Systems A and B start and stop serving packets at exactly the same times (statement (i)), all individual packet output epochs are the same and the two systems have exactly the same number of packets within them at every instant of time. ❑

III. MULTI-STAGE TREE REDUCTION

Here we use Theorem 1 to show that all multi-stage tree systems with deterministic service times can be reduced to two-stage equivalent models. Our method is to decompose the complex tree system into its "atomic" 2-stage, 2-node sub-systems, and then to apply the equivalent model result of Theorem 1 to these sub-systems.

Consider the multi-stage system of Fig. 5, which consists of nodes {1, 2, 3, 6} on the edge of the network, and nodes {4, 5, 7, 8} in the interior. Each node has an exogenous input source. Interior nodes additionally have endogenous inputs coming from previous stages.

We represent the exogenous source inputs by their normalized rates $\rho_i = \lambda_i T$ where λ_i is the arrival rate at node i and T is the service time common to all queues. By Little's Theorem ρ_i can be thought of as the loading contribution from source i. For stability, we assume that $\sum \rho_i < 1$.

Construction of an Equivalent System: Suppose we wish to understand the packet dynamics in node 8 of Fig. 5. We can reduce the complexity of the problem by applying Theorem 1 to all nodes on the edge of the network. For example, consider edge node 1 and interior node 4 as a 2-node sub-unit of the larger network. Treating the exogenous input ρ_1 as the "x" stream from Theorem 1, and the remaining inputs to node 4 as a collective "y" stream, we replace node 1 with an observation window (Fig. 6). This creates an equivalent system in which the queueing dynamics of all nodes beyond node 4 are unchanged.

Theorem 2 (Multi-Stage Tree Reduction): Suppose all queues in a multi-stage tree have deterministic service time T, and all exoge-

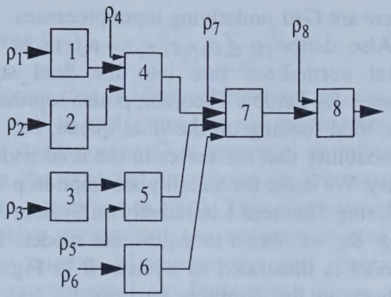


Figure 5: A 4-layer tree network.

nous input lines are stationary and independent of one another. Then, the steady-state queue length distribution for an arbitrary node of the tree is identical to the distribution in an equivalent 2-stage model.

The equivalent model is formed by removing all nodes more than 1-stage beyond the node in question, and placing their corresponding exogenous inputs directly into the remaining queues.

Example: Consider node 7 of Fig. 5. The equivalent 2-stage system model is given in Fig. 7.

Proof of Theorem 2: Consider a single node within a multi-stage tree. In order to ensure equivalent representation of this node in our simpler model, we isolate this node and keep the input processes on each line feeding into it the same. The endogenous inputs to the node in question are the departure processes from the previous stage nodes. Using Theorem 1 iteratively, we find that these departure processes are unchanged when we use observation windows to replace all nodes which are more than one stage behind the node in question. An example is shown in Fig. 6. Now, by the stationarity and independence of the lines feeding into the pre-stages, the time delays introduced by the observation windows do not effect the arrival statistics. Therefore, we can ignore all observation windows and treat these inputs as if they were directly applied to the corresponding queues of the 2-stage model. ❑

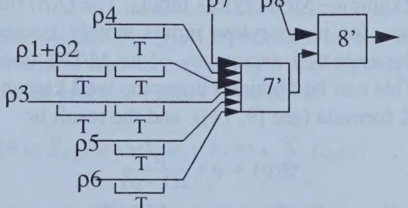


Figure 6: Using Theorem 1 iterativley to obtain an equivalent system for node 8 of Fig. 5.

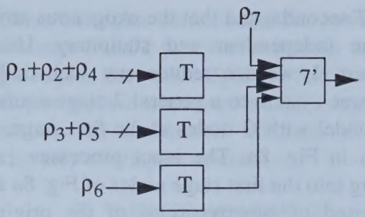


Figure 7: An equivalent model for node 7 of Fig. 5.

Theorem 2 demonstrates that solving for the occupancy statistics in a general 2-stage system (shown in Fig. 8a) provides a solution that characterizes arbitrary multi-stage configurations. For this reason, it suffices to restrict our attention to 2-stage systems.

Note that our Theorem 1 is general enough to cover the case when the deterministic service times in the downstream nodes are not identical, but are the monotonically increasing (i.e., $T_1 \le T_2 \le \ldots \le T_n$). Considering this case, we find that all of our analysis works in the exact same manner, and hence our reduction result of Theorem 2 also holds for the case when service times are monotonically increasing with each stage. However, our equivalence model fails when the deterministic service times $\{T_i\}$ decrease at some stage. Finding a good model to handle this scenario remains an open problem.

IV. EXPECTED OCCUPANCY

Theorems 1 and 2 enable exact analysis of tree networks with general types of input sources, and can be used to find queueing means, variances, and smoothing properties (see [7,8]). Here, we illustrate a simple method of calculating the expected number of packets in any node of a multi-stage tree. We assume that all service times are deterministically T seconds, and that the exogenous arrivals are independent and stationary. Using Theorem 2, we can reduce any such multi-stage tree system to a general 2-stage equivalent model with G nodes at the first stage, as shown in Fig. 8a. The input processes $\{x_i\}$ feeding into the first stage nodes of Fig. 8a are composed of superpositions of the original exogenous inputs.

Let λ_i be the arrival rate from process x_i, and let $\rho_i = \lambda_i T$ be the normalized rate. Let us suppose that the mean occupancy for a single stage queue with a superposition of these inputs can be determined. Specifically, let $Q(\rho_0 + \rho_1 + \ldots + \rho_G)$ represent the expected number of packets in a single-stage */D/1 system with inputs $x_0, x_1, \ldots x_G$. Notice that the Q function here is not simply a function of the sum total of loadings $\rho_0 + \rho_1 + \ldots + \rho_n$, but it is also a function of the actual types of processes that generate these loadings. To be more precise, we could write $Q_{X0, X1, \ldots, XG}(\rho_0 + \ldots + \rho_G)$. However, we will continue to use the more compact notation above, keeping in mind that there are G+1 underlying input processes.

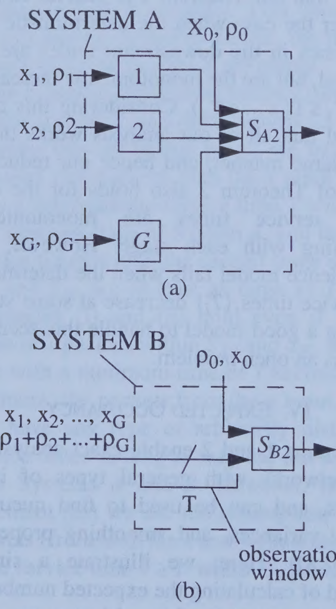


Figure 8: The general 2-stage network and its equivalent model.

Also define $\rho = \rho_0 + \rho_1 + \ldots + \rho_G$ to be the total normalized rate into the final stage queue. By Little's Theorem, ρ also represents the total loading on the final queue, i.e., the probability that the server in the final node is busy. We make the stability assumption $\rho < 1$.

Using Theorem 1 iteratively on System A of Fig. 8a, we obtain its equivalent model. This model is illustrated as System B in Fig. 8b. Observing the diagrams, we have:

$$E(\text{\# packets in System } A) = Q(\rho_1) + \ldots + Q(\rho_G) + E(\text{\# in } S_{A2}) \quad (1)$$

$$E(\text{\# packets System } B) = E(\text{\# in window}) + Q(\rho_0 + \rho_1 + \ldots + \rho_G) = (\rho - \rho_0) + Q(\rho_0 + \rho_1 + \ldots + \rho_G) \quad (2)$$

Statement (iii) of Theorem 1 indicates that the total number of packets in both Systems A and B will be the same at every instant of time. We can thus equate (1) and (2), which leads to:

$$E(\text{\# in } S_{A2}) = (\rho - \rho_0) + Q(\rho_0 + \rho_1 + \ldots + \rho_G) - (Q(\rho_1) + \ldots + Q(\rho_G)) \quad (3)$$

Equation (3) expresses the formula for expected occupancy at any node of a multi-stage tree network (with deterministic service times T) in terms of expected occupancies of single stage systems. The inputs to these single stage systems are a superposition of the exogenous inputs to the original tree network. Below we provide explicit formulas for the cases when the exogenous inputs are memoryless and when they are superpositions of independent periodic sources with period P (as in constant bit rate voice streams).

Example--Memoryless Inputs: The $Q(\rho)$ function for memoryless inputs simply becomes the expected occupancy of an M/D/1 queue. This can be obtained using the well known P-K formula (see [9, 11]), and the result is:

$$Q(\rho) = \rho + \frac{\rho^2}{2(1-\rho)} \quad (4)$$

Hence, for the system of Fig. 8a, we use (4) in (3) to find:

$$E(\#\text{ in } S_{A2}) = \rho + \frac{\rho^2}{2(1-\rho)} - \sum_{i=1}^{G} \frac{\rho_i^2}{2(1-\rho_i)} \qquad (5)$$

Notice that if identical loadings $\rho_i = \rho/G$ are used on each input line and the number of first stage queues G is increased to infinity, then the expected occupancy in the second stage converges to $Q(\rho)$. This illustrates the intuitive notion that the superposition of a large number of independent and identical low rate streams looks Poisson to a queue.

Example--Periodic (CBR) Inputs: Here we consider continuous bit rate input streams with periodic arrivals and independent phases. This type of traffic model is often used for voice data or for fixed bit rate video. There are M total flows, and each flow is a P-periodic arrival process. Stationarity holds because we assume that the first arrival of each flow is uniformly distributed over the first period interval [0, P). We have $(M - M_0)$ of the iid flows distributed over the G first stage nodes, where M_i flows feed into node i, and $M_1 + M_2 + \ldots + M_G = (M - M_0)$. The remaining M_0 flows feed directly into the final queue.

For this type of input process, the exact complementary distribution function $C_M[n]$ for single stage systems with M independent, periodic inputs of period P ($\rho = MT/P < 1$) is obtained in [10]. We state their expression: For $n \le M$, $C_M[n] =$

$$\sum_{k=1}^{M-n} \binom{M}{k+n} \left(\frac{kT}{P}\right)^{k+n} \left(1 - \frac{kT}{P}\right)^{M-k-n} \left(\frac{P/T - M + n}{P/T - k}\right) \qquad (6)$$

Using this, the expected occupancy for a single stage system can be written as:

$$Q(\rho) = Q_{P/T}[M] = \sum_{n=0}^{M-1} C_M[n] \qquad (7)$$

where we use the $Q_{P/T}[M]$ notation as a more convenient way to express Q(ρ) for this type of input process. From (3) we have:

$$E(\#\text{ in } S_{A2}) = (M - M_0)(T/P) + \sum_{n=0}^{M-1} C_M[n]$$
$$- \left[\sum_{n=0}^{M1-1} C_{M1}[n] + \ldots + \sum_{n=0}^{MG-1} C_{MG}[n] \right] \qquad (8)$$

These examples demonstrate the power that Theorems 1 and 2 provide for exactly analyzing complex networks: By these theorems, we find that node occupancy in multi-stage systems can be understood in terms of node occupancy of simpler systems.

V. A CONCAVITY RESULT

Theorem 2 indicates that multi-stage systems can be reduced to simpler 2-stage equivalent models. We thus keep our attention on 2-stage systems and explore an interesting concavity result for expected occupancy. This result provides a starting point for addressing questions concerning the optimal loading distributions on the input lines of multi-stage tree networks of queues. The result gives an important precise statement to the notion that a deterministic service time queue smooths the arrival process for downstream nodes. It shows that balanced loading minimizes the effect of this smoothing function.

Recall that $Q_{X_i}(\rho_i)$ represents the expected occupancy in a single stage queueing system with input process X_i and loading ρ_i. We wish to show a result about the convexity of $Q(\rho)$ as a function of ρ. In order for this type of statement to make sense, we must parameterize an input stream by its loading value. The idea is to consider only input processes that can be represented as a sum of iid subflow processes. The loading is then determined by the number of subflow components used.

Definition: A process X with loading ρ is *M-distributable* iff:

$$X = \sum_{i=1}^{M} X_i \qquad (9)$$

where each of the component processes X_i is iid with loading $\rho_i = \rho/M$.

A process is *infinitely distributable* if it is M-distributable for all M.

The Poisson process is the canonical example an infinitely distributable process although it is not the only one.

For an M-distributable input process we can now consider the function $Q(k\Delta)$ for

$\Delta = \rho / M$ and $k = 0, 1, \ldots, M$. Note that such a $Q(k\Delta)$ function is monotonically increasing in k as adding another subflow to the input flow can only increase the number in the queueing system. Finally, we note that we must define what we mean by a convex function with a discrete domain.

Definition: The function $Q(k\Delta)$ is *convex* in k iff:

$$Q((k+1)\Delta) - Q(k\Delta) \geq Q(k\Delta) - Q((k-1)\Delta)$$

for $k = 1, 2, \ldots, M-1$.

This definition of convexity preserves the usual notions of convexity generally known for functions of a continuous variable or continuous multivariables.

Theorem 3 (Convexity of average queue length): For an M-distributable input process entering a deterministic service time queue, the function giving the expected number of packets in the queueing system, $Q(k\Delta)$, is convex.

Proof of Theorem 3: Consider the deterministic service time queues represented in Fig. 9. A switch has been added to the top input as we will compare the situation in system S_2 when the top input is present and when it is not. Let $\bar{n}_{ON}(S_2)$ ($\bar{n}_{OFF}(S_2)$) be the average number of packets in the S_2 queueing system when the top source is ON (OFF). Using Theorem 1 we can write: (eqs. 10,11)

$$\bar{n}_{ON}(S_2) + Q(k\Delta) + Q(\Delta) = (k+1)\Delta + Q((k+1)\Delta)$$

$$\bar{n}_{OFF}(S_2) + Q((k-1)\Delta) + Q(\Delta) = k\Delta + Q(k\Delta)$$

which lead to eqs. (12, 13) below:

$$\bar{n}_{ON}(S_2) - (k+1)\Delta = Q((k+1)\Delta) - Q(k\Delta) - Q(\Delta)$$

$$\bar{n}_{OFF}(S_2) - k\Delta = Q(k\Delta) - Q((k-1)\Delta) - Q(\Delta).$$

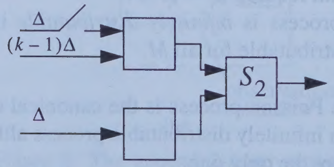


Figure 9: The queueing system used for the convexity proof, with an ON/OFF switch.

Consider the queueing system S_2 when the top source is ON. While the expected number in the S_2 system is $\bar{n}_{ON}(S_2)$, the expected number in the S_2 server alone is equal to the loading of the S_2 system, $(k+1)\Delta$. The left side of (12) represents the expected number in the *queue* of system S_2 with the top source ON while the left side of (13) represents the expected number in the *queue* of system S_2 with the top source OFF. A sample function argument similar to that used in the proof of Theorem 1 shows formally the not surprising fact that the average number of packets in the *queue* of system S_2 is never less when the top source is ON than when the top source is OFF:

$$\bar{n}_{ON}(S_2) - (k+1)\Delta \geq \bar{n}_{OFF}(S_2) - k\Delta \quad (14)$$

So,

$$Q((k+1)\Delta) - Q(k\Delta) \geq Q(k\Delta) - Q((k-1)\Delta) \quad (15)$$

❑

This theorem also holds for infinitely-distributable inputs. Analysis shows that as $M \to \infty$, the $Q(\rho)$ curve becomes a continuous, monotonically increasing convex function of a continuous variable. Memoryless inputs are infinitely-distributable, and we clearly see that the corresponding $Q(\rho)$ function given in (4) is continuous, monotonically increasing, and convex.

Two Stage Systems: We now consider the two stage system with distributable inputs shown in Fig. 10. We are interested in finding an "optimal" distribution for the input loadings. Should all loadings be distributed over the pre-stage units equally? Will an imbalance be helpful or hurtful? The following theorem provides insight into this question. We get the proof of the theorem as a straightforward application of Theorem 3.

Theorem 4 (Convexity and concavity of two stage systems): Assume the input process $(\rho_1, \ldots, \rho_G)$ in Fig. 10 is distributable with fixed total loading $\rho = \rho_1 + \ldots + \rho_G$. Then:

(i) The expected occupancy in the first stage system of Fig. 10 is a convex symmetric function of the input loadings $\rho_1, \ldots, \rho_G$. The occupancy at this stage is minimized when all loadings are distributed uniformly.

(ii) The expected occupancy in the final stage of the system in Fig. 10 is a concave symmetric function of the input loadings $\rho_1, \ldots, \rho_G$. The occupancy at this stage is maximized when all loadings are distributed uniformly.

Proof of Theorem 4:

Proof of (i): The expected occupancy at the first stage is just the sum of the expected occupancies in the individual first stage queues: $Q(\rho_1) + Q(\rho_2) + \ldots + Q(\rho_G)$. Each of the individual terms of the sum are convex functions by Theorem 3. Hence, the sum will also be convex. Clearly, the expected occupancy at the first stage is a symmetric function of the input loadings. The minimum of this convex symmetric function is thus achieved when all input loadings are distributed equally. ❑

Proof of (ii): Using Theorem 1, we can use an equivalent model of the system in Fig. 10 to find that the expected occupancy in the whole system is $\rho + Q(\rho)$, where $\rho = \rho_1 + \ldots + \rho_G$. Note that this value depends only on the total loading, not on the particular loading assignments $(\rho_1, \ldots, \rho_G)$. Hence, the expected occupancy in the second stage is just the total occupancy minus the first stage occupancy:

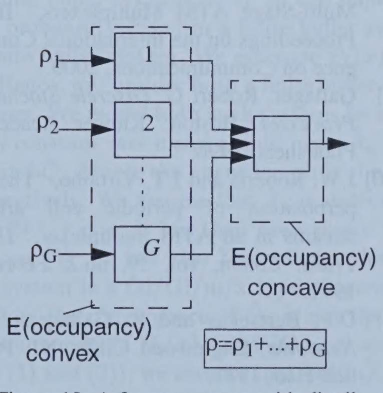


Figure 10: A 2-stage system with distributable inputs.

$$E(\text{Occupancy at second stage}) = \rho + Q(\rho) - [Q(\rho_1) + \ldots + Q(\rho_G)] \quad (16)$$

For a fixed total loading ρ, the above is just a constant minus a convex symmetric function. Hence, the expected second stage occupancy is a concave symmetric function, whose maximum is achieved when all individual loadings ρ_i are equal ($\rho_i = \rho/G$). ❑

Notice that adding an exogenous input ρ_0 to the second stage in Fig. 10 (and redefining the total loading to be $\rho = \rho_0 + \rho_1 + \ldots + \rho_G$) does not change the concavity property of the expected occupancy in this node as a function of $(\rho_0, \rho_1, \ldots, \rho_G)$. However, it is clear that the symmetry is then only with respect to modifications of the original variables $\rho_1, \ldots, \rho_G$.

Theorem 4 provides insight into how changing the input loading distribution will affect the overall system of Fig. 10. In particular, the theorem gives a statement of how deterministic queues smooth the arrival process for future stages. When most of the traffic goes through only a few first stage queues, there is more smoothing than when the traffic is evenly spread over many first stage queues. Our theorem is in terms of expected occupancies, but note that these values are directly related to the amount of buffer slots we require for a given packet loss threshold ε. Smoothing properties of these networks are further explored in [8].

VI. CONCLUSIONS

Packet occupancy in multi-stage tree networks of identical deterministic service time queues can be tractably analyzed using the equivalent model results of Theorems 1 and 2. We have shown that if the exogenous inputs to such networks are independent and stationary, then multi-stage analysis can be reduced to analysis of a simpler 2-stage system. The inputs to the 2-stage system are simply superpositions of the original exogenous input processes.

Using this theory, we have developed an expression for the mean occupancy in any node of a multi-stage tree network in terms of the mean occupancy of a single stage (one queue) system with an arrival process consisting of a superposition of the original exogenous inputs. As examples, this expression was written explicitly for the cases when the exogenous inputs are (i) memoryless, and (ii) periodic.

Focusing our attention on 2-stage systems, we have further explored convexity behavior of expected node occupancies. We have created a definition of a "distributable input," and have shown that simple, comparative observations about 2-stage systems and their equivalent models lead to several convexity/concavity results about expected packet occupancy in the first and second stages of our system. Specifically, we have shown that the expected occupancy function $Q(\rho)$ for a single queue is a convex, monotonically increasing function of the loading ρ. This implies that the number of packets in the first stage is a convex symmetric function of the exogenous input rates. Likewise, the second stage occupancy is a concave symmetric function of the exogenous input rates. For the second stage, the maximum of this concave symmetric function is achieved by the uniform rate distribution.

Expanding the 2-stage result via Theorem 2 implies that expected occupancy is concave in the rate assignment for any interior node of a multi-stage tree system. Nodes on the edge of the network have convex expected occupancies. These results provide much insight into tree systems and the smoothing properties of deterministic service time queues. It is interesting to pursue extensions of this theory towards the development of buffer requirements and optimal loading assignments in ATM networks with various types of data traffic.

Notice that our results are stated in terms of fixed length packets flowing through the network, and these packets have identical deterministic service times T in all network nodes. The exact same results apply to deterministic service time queues with monotonically increasing service times $T_1 \le T_2 \le \ldots \le T_n$. However, decreasing service times pose a formidable challenge. It would be interesting to develop models to cover this type of network.

REFERENCES:

[1] M. Shalmon and M.A. Kaplan, "A tandem network of queues with deterministic service and intermediate arrivals," *Operations Research*, vol. 32, no. 4, July-August 1984.

[2] J. Morrison, "Two discrete-time queues in tandem," *IEEE Trans. Comm.*, vol. COM-27, no. 3 March 1979.

[3] S.K. Walley and A.M. Viterbi, "A tandem of discrete-time queues with arrivals and departures at each stage," *Queueing Systems*, vol. 23, pp. 157-176, 1996.

[4] O.J. Boxma, "On a tandem queueing model with identical service times at both counters, I and II," *Adv. Appl. Prob.*, vol. 11, pp. 616-643, 644-659, 1979.

[5] E. Modiano, J.E. Wieselthier, A. Ephremides, "A simple analysis of average queueing delay in tree networks," *IEEE Trans. Info. Th.*, vol. IT-42, no. 2, pp. 660-664. March 1996.

[6] J.A. Morrison, "A combinatorial lemma and its application to concentrating trees of discrete-time queues," *Bell Sys. Tech. J.*, vol. 57, no.5, pp. 1645-1652, May-June 1978.

[7] M.J. Neely with C.E. Rohrs, *Queue Occupancy in Single-Server, Deterministic Service Time Tree Networks*. Masters Thesis, MIT LIDS, March 1999.

[8] M.J. Neely and C.E. Rohrs, "Inequality Comparisons and Traffic Smoothing in Multi-Stage ATM Multiplexers," IEEE Proceedings on the International Conference on Communications, 2000.

[9] Gallager, Robert G. *Discrete Stochastic Processes*. Boston: Kluwer Academic Publishers, 1996.

[10] J.W. Roberts and J.T. Virtamo, "The superposition of periodic cell arrival streams in an ATM multiplexer," *IEEE Trans. Comm*, vol. 39, no.2. February 1991.

[11] D.P. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1987.

Estimating the Loss Probability for Finite Buffer Multiplexers

Han S. Kim and Ness B. Shroff*

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285, U.S.A.
{hkim, shroff}@ecn.purdue.edu

Abstract

In this paper we propose an approximation for the loss probability in a finite buffer system by making a simple mapping from the tail probability in an infinite buffer system. We focus on the case of a high-speed network where a large number of sources are expected to be multiplexed so that we model the input process as a Gaussian process. We also provide a strong asymptotic relationship between our approximation and the actual loss probability for a fairly large class of Gaussian input processes. We validate our approximation for a variety of well established traffic models (such as self-similar models, Markov modulated models, etc.) and also trace-driven simulations. We find that our approximation provides a very good estimate to the loss probability over all tested buffer sizes.

1 Introduction

The loss probability is one of the most important QoS (Quality of Service) measures in communication networks. But there have been relatively few works on the loss probability in finite buffer systems [1, 2]. In this paper, we propose a simple method to estimate the loss probability $P_L(x)$ in a finite buffer system from the tail probability $\mathbb{P}\{Q > x\}$ of an infinite buffer system. We estimate $P_L(x)$ by making a simple mapping from $\mathbb{P}\{Q > x\}$.

Hence, we consider both a finite buffer queueing system and an infinite buffer queueing system. We model both such systems by a discrete-time fluid queue consisting of a server with constant rate c and a fluid input λ_n. Both queues are fed with the same input. Let $\hat{Q}_n$ and Q_n denote the queue length in the finite queue and in the infinite queue at time n, respectively. We assume that λ_n is stationary and ergodic and that the system is stable, i.e., $\mathbb{E}\{\lambda_n\} < c$. It has then been shown that Q_n converges to a stationary and ergodic process [3]. It has also been shown that $\hat{Q}_n$ converges to a stationary process when the system is a GI/GI/m/x type of queue [4], or a G/M/m/x type of queue [5]. Since proving the convergence of $\hat{Q}_n$ is not the focus of this paper, and moreover, practically measurements of $P_L(x)$ and $\mathbb{P}\{Q > x\}$ are based on "time average" assuming ergodicity (see (1) and (2)), we assume that both $\hat{Q}_n$ and Q_n started at $n = -\infty$, and that they are ergodic and stationary.[1] The time index n is often omitted to represent the stationary distribution, i.e., $\mathbb{P}\{Q > x\} = \mathbb{P}\{Q_n > x\}$ and $\mathbb{E}\{\lambda\} = \mathbb{E}\{\lambda_n\}$.

---

*Please address all correspondence to Ness B. Shroff

The loss probability, $P_L(x)$, for a buffer size x is defined as the long-term ratio of the amount of fluid lost to the amount of fluid fed. It is expressed as

$$P_L(x) = \lim_{N\to\infty} \frac{\sum_{k=1}^{N}(\hat{Q}_{k-1} + \lambda_k - c - x)^+}{\sum_{k=1}^{N} \lambda_k} = \frac{\mathbb{E}\{(\hat{Q}_{n-1} + \lambda_n - c - x)^+\}}{\mathbb{E}\{\lambda_n\}}, \tag{1}$$

where $(x)^+$ denotes $\max\{x, 0\}$, and the second equality is due to the ergodicity assumption. The tail probability $\mathbb{P}\{Q > x\}$ is defined as the amount of time the fluid in the infinite buffer system spends above level x divided by the total time. It is expressed as:

$$\mathbb{P}\{Q > x\} \;=\; \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} I(Q_k > x), \tag{2}$$

where $I(A) = 1$ if A is true; $I(A) = 0$ otherwise. From now on, when we write "loss probability" it will only be in the context of a finite buffer system, and when we write "tail probability" it will only be in the context of an infinite buffer system.

Since $\mathbb{P}\{Q > x\}$ is averaged by time, and $P_L(x)$ is averaged by the input, in general there is no relationship between these two quantities. However, very often, $P_L(x)$ is approximated as $P_L(x) \approx \mathbb{P}\{Q > x\}$. Empirically, it has been found that this approximation usually provides an upper bound (sometimes a very poor bound) to the loss probability, although in general this cannot be proven. Based on extensive simulation experience, we have found that the curves, $\log P_L(x)$ and $\log \mathbb{P}\{Q > x\}$ versus x, exhibit a very similar shape, i.e., one looks like a parallel shift of the other. Further, it has been shown in [7] that for M/Subexponential/1 and GI/Regularly-varing/1 with i.i.d. interarrival times and i.i.d. service times, $\mathbb{P}\{Q > x\}/P_L(x)$ converges to a constant, as $x \to \infty$. Hence, it seems reasonable that if we have a good estimate of $\mathbb{P}\{Q > x\}$ for all x and a way to calculate $P_L(x)$ for one point of $x = 0$, then we can calculate $P_L(x)$ for all x. *This is the basic idea that drives this paper.* We will also show that asymptotically the loss probability and the tail probability curves are quite similar, and if they diverge, they do so very slowly. This provides theoretical justification for our approximation.

For our study, we focus on the case when the aggregate traffic can be characterized by a Gaussian process. Gaussian processes have received significant attention as good models for the arrival process to a high-speed multiplexer [8, 9]. There are many reasons for this. Due to the huge link capacity of high-speed networks, hundreds or even thousands of network applications are likely to be served by a network multiplexer. Also, when a large number of sources are multiplexed, characterizing the input process with traditional Markovian models results in computational infeasibility problems [2] that are not encountered for Gaussian processes. Finally, recent network traffic studies suggest that certain types of network traffic may exhibit self-similar type of long-range dependence, and various Gaussian processes can be used to model this type of behavior.

This paper is organized as follows. In Section 2, we review the maximum variance asymptotic (MVA) results for the tail probability, and then demonstrate how to obtain similar results for the loss probability. In Section 3, we compare our approach to an approach based on the *many sources asymptotics*. In Section 4, we validate our result with several numerical examples. In Section 5, we find the asymptotic relationship between the loss probability and our approximation. We finally state the conclusions in Section 6. In the text we do not provide proofs of theorems due to spacing problem. All proofs can be found in [10].

---

[1] In [6], we have studied the relationship without assuming ergodicity of $\hat{Q}_n$.

2 MVA Approximation for Loss

The first component in our development of an approximation for $P_L(x)$ is to find a good estimate of $\mathbb{P}\{Q > x\}$. Fortunately, this part of the problem has already been solved in [8]. By developing results based on Extreme Value Theory, it has been found that the Maximum Variance Asymptotic (MVA) approach (first named in [8]) provides a very accurate estimate of the tail probability. Here we review it very briefly.

As mentioned before, we focus on the case when the aggregate traffic can be characterized by a Gaussian process, hence the input process λ_n is assumed to be a stationary and ergodic Gaussian process. Let $\bar{\lambda} := \mathbb{E}\{\lambda_n\}$ and $\sigma^2 :=\text{Var}\{\lambda_n\}$. We define a stochastic process X_n as

$$X_n := \sum_{k=1}^{n} \lambda_k - cn. \tag{3}$$

Then, it has been shown in [3] that $\mathbb{P}\{Q > x\} = \mathbb{P}\{\sup_{n\geq 1} X_n > x\}$. Let $C_\lambda(l)$ be the autocovariance function of λ_n. Then, the variance of X_n can be expressed in terms of $C_\lambda(l)$. For each $x > 0$, define the normalized variance $\sigma_{x,n}^2$ of X_n as

$$\sigma_{x,n}^2 := \left\{ \frac{\text{Var}\{X_n\}}{(x - \mathbb{E}\{X_n\})^2} \right\} = \left\{ \frac{nC_\lambda(0) + 2\sum_{l=1}^{n-1}(n-l)C_\lambda(l)}{(x+\kappa n)^2} \right\}, \tag{4}$$

where $\kappa := c - \bar{\lambda}$. Let m_x be the reciprocal of the maximum of $\sigma_{x,n}^2$ for given x, i.e.,

$$m_x := \frac{1}{\max_{n\geq 1} \sigma_{x,n}^2} = \min_{n\geq 1} \frac{(x+\kappa n)^2}{\text{Var}\{X_n\}}. \tag{5}$$

In [8], under a set of conditions on $C_\lambda(l)$, it has been shown that $e^{-\frac{m_x}{2}}$ is an asymptotic upper bound on the tail probability. Further the following global lower bound has also been examined in [8]:

$$\mathbb{P}\{Q > x\} = \mathbb{P}\{\sup_{n\geq 1} X_n > x\} \geq \sup_{n\geq 1} \mathbb{P}\{X_n > x\} = \mathbb{P}\{X_{n_x} > x\} = \Psi(\sqrt{m_x}), \tag{6}$$

where $\Psi(z) := \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-\frac{y^2}{2}} dy$ is the tail of the standard Gaussian distribution, and n_x is the time index at which the normalized variance of X_n is maximized, i.e., m_x is minimized. After extensive simulation studies we have found that the estimate $e^{-\frac{m_x}{2}}$ called the Maximum Variance Asymptotic (MVA) approximation, behaves as if it were a global upper bound. Moreover, the difference between this asymptotic upper bound and the global lower bound has a minimum at $x = 0$. Hence, the value of $e^{-\frac{m_x}{2}}$ at $x = 0$ is expected to be very close to the actual value. For a given x both n_x and m_x cannot generally be obtained in a simple closed form, but they can be found by simple local search algorithms within a small number of iterations. are likely to be used to evaluate them. Furthermore, n_x may not be unique especially for a small value of x. When $x = 0$, however, we can obtain them right away as demonstrated in the following proposition.

Proposition 1 *Let n_x be the value of n at which $\sigma^2_{x,n}$ attains its maximum. Then, $n_0 = 1$ and*

$$m_0 = \frac{\kappa^2}{C_\lambda(0)}. \tag{7}$$

□

Now, we need to calculate $P_L(0)$. Since λ_n is Gaussian, the mean and the variance provide sufficient information to calculate $P_L(0)$, i.e.,

$$P_L(0) = \frac{\mathbb{E}\{(\lambda_n - c)^+\}}{\mathbb{E}\{\lambda_n\}} = \frac{1}{\bar{\lambda}} \int_c^\infty (r-c) e^{-\frac{(r-\bar{\lambda})^2}{2\sigma^2}} \, dr \tag{8}$$

As long as the number of input sources is large enough for the aggregate traffic to be characterized as a Gaussian process, (8) gives a very accurate estimate and it is often called the Gaussian approximation [11]. Note that $C_\lambda(0) = \sigma^2$ and $\kappa = c - \bar{\lambda}$ in (7). From (7) and (8), we have

$$P_L(x) \approx \alpha e^{-\frac{m_x}{2}}, \tag{9}$$

where

$$\alpha = \frac{1}{\bar{\lambda}} \exp\left(\frac{(c-\bar{\lambda})^2}{2\sigma^2}\right) \int_c^\infty (r-c) e^{-\frac{(r-\bar{\lambda})^2}{2\sigma^2}} \, dr.$$

We call this above approximation the *MVA approximation for loss* (abbreviated as "MVA-Loss"). The MVA approach is based on the *large buffer asymptotics* and it also applies in the context of the *many sources asymptotics* [12]. We next compare this approach with an approximation based on the many sources asymptotics.

3 Comparison with the L-M Result

The many sources asymptotics has been widely studied and can be found in many papers on queueing analysis using large-deviation technique [13, 1, 14]. Most of the papers deal with the tail probability rather than the loss probability. In [1], the authors developed the first result on the loss probability based on the many sources asymptotics, which we call the Likhanov-Mazumdar approximation for loss (abbreviated as "L-M"). This result was obtained for a fairly general class of arrival processes and is much stronger than typical large-deviation types of results.

Consider N i.i.d. sources, each with input rate $\lambda_n^{(i)}$, $n \in \{0, 1, 2, \ldots\}, i \in \{1, 2, \ldots, N\}$. It is assumed that the moment generating function of $\lambda_n^{(1)}$ exists, and that the input rate $\lambda_n^{(i)}$ is bounded. Let $\bar{\lambda}_1 = \mathbb{E}\{\lambda_n^{(1)}\}$, $\phi_n(\theta) = \mathbb{E}\{e^{\theta \sum_{k=1}^n \lambda_k^{(1)}}\}$, and N, C, B be the number of sources, the link capacity per source, and the buffer size per source, respectively. The L-M approximation has the following form (Equation (2.6) in [1]):

$$P_L(NB) \approx \frac{e^{-N I_{\hat{n}}(C,B)}}{\theta_{\hat{n}}^2 \bar{\lambda}_1 \sqrt{2\pi\sigma_{\hat{n}}^2 N^3}}, \tag{10}$$

where

- θ_n is the unique solution to $\frac{\phi_n'(\theta_n)}{\phi_n(\theta_n)} = Cn + B$,

- $\sigma_n^2 = \frac{\phi_n''(\theta_n)}{\phi_n(\theta_n)} - (Cn + B)^2$,
- $I_n(C, B) = (Cn + B)\theta_n - \log \phi_n(\theta_n)$,
- $\hat{n}$ is a value of n such that $I_n(C, B)$ is minimized.

Consider the numerical complexity of (10). Suppose that we calculate (10) for given N, C, B, and $\lambda_n^{(1)}$. In general, since there are no closed-form solutions for θ_n and $\hat{n}$, we have to find them numerically. Two iteration loops are nested; The inner loop iterates over θ to find θ_n for given n, and the outer loop iterates over n to find $\hat{n}$. Hence, it can take a very long time to find a solution of (10) by numerical iteration, especially when the buffer size B is large. Here we have compared the numerical complexity, and in the next section we will investiagte the accuracy via simulations.

4 Numerical Validation of MVA-Loss

In this section, we investigate the accuracy of the proposed method by comparing our technique with simulation results. To improve the reliability of the simulation, we use *Importance Sampling* whenever applicable. Due to spacing problem, only a few examples are provided here. More results can be found in [10].

4.1 Simulation with Gaussian Input

In this example, we apply L-M even though Gaussian examples violate the assumption of boundedness of the input rate, since after all the simulated Gaussian processes have bounded rates. As mentioned, it takes a very long time to find a solution of (10) by numerical iteration especially when the buffer size B is large. But for the Gaussian case, we have a closed-form solution for θ_n, i.e., as $\phi_n(\theta) = e^{\theta m(n) + \frac{1}{2}\theta^2 v(n)}$ with $m(n) = \mathbb{E}\{\sum_{k=1}^{n} \lambda_k^{(1)}\}$ and $v(n) = \mathrm{Var}\{\sum_{k=1}^{n} \lambda_k^{(1)}\}$, we have $\theta_n = (Cn + B - m(n))/v(n)$. Hence for given C and B both $I_n(C, B)$ and $\sigma_{NB,n}^2$ (the normalized variance of X_n) are expressed in terms of $n, m(n)$, and $v(n)$, the complexity to find $\hat{n}$ in L-M now requires the same effort as the complexity to find n_{NB} in MVA-Loss. We call this approach the L-M:Gaussian approximation, and abbreviate as "L-M:Gaussian".

This example deals with a very well known input process, the fractional Brownian motion (FBM) process, which is the classical example of a self-similar process [9].[2] The results are shown in Fig. 1, demonstrating the accuracy of MVA-Loss, even for self-similar sources. Due to the difficulty in applying importance sampling techniques to obtain loss probabilities for self-similar traffic, in Fig. 1, we show probabilities only as low as 10^{-6}. Note that the input traffic is characterized by a superposition of FBM processes with different Hurst parameters, which is more realistic to accommodate the heterogeneity of sources that high-speed networks will carry. It seems that the error in L-M becomes larger as the utilization becomes higher. However, MVA-Loss works well even in this regime. More results for different types of Gaussian processes in [10] also show a very close match over the entire range of buffers tested.

---

[2] Since continuous-time FBM processes cannot be simulated, one typically uses a discrete-time version, called fractional Gaussian noise (FGN), with the autocovariance function given by $C_\lambda(l) = \frac{\sigma^2}{2}(|l-1|^{2H} + |l+1|^{2H} - 2|l|^{2H})$, where $H \in [0.5, 1)$ is the Hurst parameter.

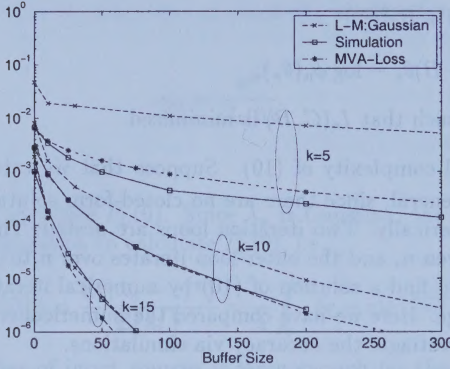


Figure 1: The exact loss probability, MVA-Loss and L-M for the superposition of two FBM processes. (Hurst paramters: $H_1 = 0.6, H_2 = 0.8$; mean rates: $\bar{\lambda}_1 = 150$, $\bar{\lambda}_2 = 150$; variances: $\sigma_1^2 = 50, \sigma_2^2 = 50$; service rate - mean rate: $\kappa = 5, 10, 15$)

4.2 Simulation with Non-Gaussian Input

Consider a very practically important case of multiplexed Markov Modulated Fluid (MMF) ON-OFF sources. The input MMF process has widely been used to model voice traffic source. In this example, transition matrix is [0.9833, 0.0167; 0.025, 0.975] and the rate vector is [0 cells/slot, 0.85 cells/slot]. These values are chosen for a 45 Mbps ATM link with 10 msec time slot and 53 byte ATM cell. Since MVA-Loss assumes that the input process is Gaussian, we regard the aggregate of non-Gaussian inputs as a Gaussian having the same mean and autocovariance with the original, i.e., mean = 987.18 and $C_\lambda(l) = 503 \times 0.958^{|l|}$. In Fig. 2, the result is compared to existing methods such as Effective Bandwidth and Hybrid scheme [2, 15]. Fig. 2 also shows two types of L-M curves. The one labeled "L-M:MMF" is obtained from the exact moment generating function of the MMF source and the other labeled "L-M:Gaussian" is obtained from that of the Gaussian source having the same mean and autocovariance. As shown in Fig. 2, MVA-Loss obtains the loss probability calculations accurately and better than the other techniques.

We next investigate the accuracy of our approximation when the sources to the queue are generated from actual MPEG video traces. The trace used to generate this simulation result comes from an MPEG-encoded action movie (007 series) which has been found to exhibit long-range dependence. In Fig. 3, 240 MPEG sources are multiplexed and served at 3667 cells/slot ($OC-3$ line), where we assume 25 frames/sec and a 10 msec slot size. MVA-Loss still works better than others.

5 Theoretical Validation of MVA-Loss

To provide theoretical justification for our approximation, we find a strong asymptotic relationship between the loss probability and the tail probability. More specifically, under

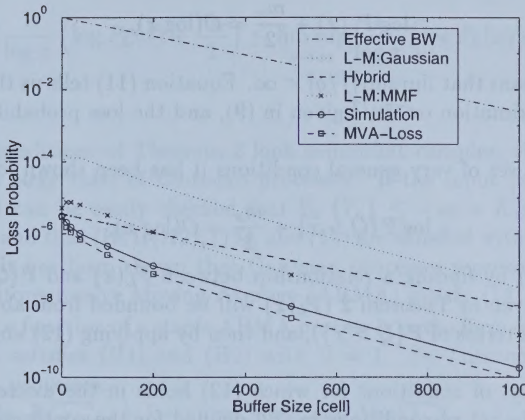


Figure 2: The loss probability curves for 2900 voice sources. (transition matrix = [0.9833, 0.0167; 0.025, 0.975]; rate vector = [0 cells/slot, 0.85 cells/slot]; mean rate (total): $\bar{\lambda} =$ 987.18 cells/slot; service rate - mean rate: $\kappa = 73.82$ cells/slot)

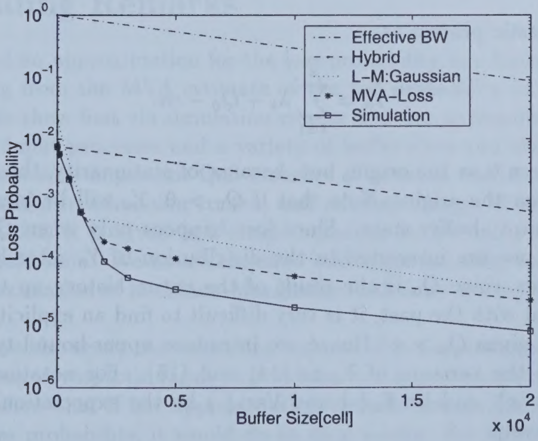


Figure 3: The loss probability curves for 240 real MPEG traces from a 007 movie. (mean rate (total): $\bar{\lambda} = 3469.45$ cells/slot; service rate - mean rate: $\kappa = 197.55$ cells/slot)

some conditions (to be defined later in Theorem 2), we find that

$$\log P_L(x) + \frac{m_x}{2} = O(\log x), \tag{11}$$

where $f = O(g)$ means that $\limsup |f/g| < \infty$. Equation (11) tells us that the divergence between the approximation $\alpha e^{-m_x/2}$ given in (9), and the loss probability is very slow if at all.

In [12], under a set of very general conditions it has been shown for the continuous-time case that

$$\log \mathbb{P}\{Q > x\} + \frac{m_x}{2} = O(\log x). \tag{12}$$

We can obtain (11) by finding a relationship between $P_L(x)$ and $\mathbb{P}\{Q > x\}$ under the set of conditions given in Theorem 2 ($P_L(x)$ will be bounded from above and below by some expressions in terms of $\mathbb{P}\{Q > x\}$), and then by applying (12) and some properties of m_x.

We first list a set of conditions for which (12) holds in the discrete-time case that are equivalent to the set of conditions in [12] defined for the continuous-time case. Let $v_n :=\mathrm{Var}\{X_n\}$, $\psi(n) := \log v_n$, and $\beta := \lim_{n\to\infty} \psi(n)/\log n$ (assuming that the limit exists).

$$\text{(H1)} \qquad \lim_{n\to\infty} n\,[\psi(n+1) - \psi(n)] = \beta.$$

$$\text{(H2)} \qquad v_n \overset{n\to\infty}{\sim} Sn^{\beta} \text{ for some } S > 0.$$

The notation $f(n) \overset{n\to\infty}{\sim} g(n)$ means that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. The parameter β cannot be larger than 2 due to the stationarity of λ_n, and $\beta \in (0, 2)$ covers the majority of non-trivial stationary Gaussian processes. The Hurst parameter H is related to β by $\beta = 2H$.

Define a stochastic process as

$$Y_n = \sum_{k=1}^{n} \lambda_k + Q_0 - cn. \tag{13}$$

Here we have chosen 0 as the origin, but, because of stationarity, the distribution of Y_n does not depend on the origin. Note that if $Q_0 > 0$, Y_n will be identical to Q_n until Q_n reaches the empty-buffer state. Since loss happens only when Q_n is greater than the buffer size x, we are interested in the distribution of Y_n after loss happens (i.e., $Q_0 > x$). However, since Q_0 is the result of the entire history up to time 0 and the future is correlated with the past, it is very difficult to find an explicit expression of the distribution of Y_n given $Q_0 > x$. Hence, we introduce upper-bound types of conditions on the mean and the variance of Y_n as (14) and (15). For notational simplicity, let $\mathbb{P}_x\{\cdot\} = \mathbb{P}\{\cdot|Q_0 > x\}$, and let $\mathbb{E}_x\{\cdot\}$ and $\mathrm{Var}_x\{\cdot\}$ be the expectation and the variance under $\mathbb{P}_x$, respectively.

We now state our main theorem omitting the proof due to the space pr.

Theorem 2 *Assume conditions (H1) and (H2). Further assume that for any* $\epsilon > 0$, *there exist* $x_0, K, M,$ *and* α *such that*

$$\mathbb{E}_x\{Y_n\} \;\le\; (-\kappa + \epsilon)n, \tag{14}$$

$$\mathrm{Var}_x\{Y_n\} \;\le\; Kn^{\beta}, \tag{15}$$

for all $x \geq x_0$ *and* $n \geq Mx^{\alpha}$. *Then,*

$$-\infty < \liminf_{x\to\infty} \frac{1}{\log x}\left(\log P_L(x) + \frac{m_x}{2}\right) \leq \limsup_{x\to\infty} \frac{1}{\log x}\left(\log P_L(x) + \frac{m_x}{2}\right) < \infty. \qquad (16)$$

□

Though the conditions of Theorem 2 look somewhat complex, they are expected to be satisfied by a large class of Gaussian processes. If the input process is i.i.d. with $C_\lambda(l) = S\delta(l)$, it can be easily checked that $\mathbb{E}_x\{Y_n\} \leq -\kappa n + K_1 x$ for some K_1, and $\mathrm{Var}_x\{Y_n\} = Sn$, and that (H1), (H2), (14), and (15) are satisfied with $\beta = 1, K = S, M = \frac{K_1}{\epsilon}$, and $\alpha = 1$. It has been shown that Gaussian processes represented by the form of finite-ordered Autoregressive Moving Average (ARMA) satisfy (14) and (15) [6]. Since the autocovariance function of a stable ARMA process is in the form of $C_\lambda(l) = \sum_{i=1}^{N} a_i p_i^{|l|}$ with $|p_i| < 1$, it satisfies (H1) and (H2) with $\beta = 1$. So Theorem 2 is applicable to Gaussian ARMA processes.

More generally, $\mathbb{E}\{\sum_{k=1}^{n} \lambda_k\} - cn = -\kappa n$, and $\mathrm{Var}\{\sum_{k=1}^{n} \lambda_k\} \overset{n\to\infty}{\sim} Sn^{\beta}$ under (H1) and (H2). Thus, for each x, $\mathbb{E}_x\{Y_n\} \overset{n\to\infty}{\sim} -\kappa n$ and $\mathrm{Var}_x\{Y_n\} \overset{n\to\infty}{\sim} Sn^{\beta}$, and we can find $K = K(\epsilon, x)$, $M = M(\epsilon, x)$, and $\alpha = \alpha(\epsilon, x)$ as small as possible. If $\sup_x K(\epsilon, x)$, $\sup_x M(\epsilon, x)$, and $\sup_x \alpha(\epsilon, x)$ are finite, then (14) and (15) hold. We conjecture that they are all finite for a large class of stationary Gaussian processes, and we are trying to show it.

Note that the rightmost inequality (limsup part) in (16) holds without conditions (14) and (15), and it agrees with empirical observations that the tail probability curve provides an upper bound to the loss probability curve.

6 Concluding Remarks

We have proposed an approximation for the loss probability in a finite queue by making a simple mapping from the MVA estimate of the tail probability in the corresponding infinite queue. We show first via simulation results that our approximation is very accurate for different input processes and a variety of buffer sizes and utilization. Since the loss probability is a very important QoS measure of network traffic, this approximation will serve to be useful in admission control and network design. Another feature of the approximation is that it is given in a single equation format and hence can easily be implemented in real-time. We have compared our approximation to the many sources asymptotic approximation of Likhanov-Mazumdar both from a complexity and accuracy point of view.

In this paper we also study the theoretical aspects of our approximation. In particular, we provide a strong asymptotic result that relates our approximation to the actual loss probability. We show that if our approximation were to diverge (with increasing buffer size) from the loss probability, it would do so very slowly. For future work we plan on simplifying the conditions given in Theorem 2 and to extend the approximation result to a network of queues.

References

[1] N. Likhanov and R. R. Mazumdar, “Cell-loss asymptotics in buffers fed with a large

number of independent stationary sources," in *Proceedings of IEEE INFOCOM*, San Francisco, CA, 1998.

[2] N. B. Shroff and M. Schwartz, "Improved Loss Calculations at an ATM Multiplexer," *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 411–422, Aug. 1998.

[3] R. M. Loynes, "The Stability of a Queue with Non-independent Inter-arrival and Service Times," *Proc. Cambridge Philos. Soc.*, vol. 58, pp. 497–520, 1962.

[4] D.P. Heyman and W. Whitt, "Limits for Queues as the Waiting Room Grows," *Queueing Systems*, vol. 5, pp. 381–392, 1989.

[5] A. Brandt, P. Franken, and B. Lisek, *Stationary Stochastic Models*, John Wiley & Son, New York, 1990.

[6] H. S. Kim and N. B. Shroff, "On the Asymptotic Relationship between the Overflow Probability in an Infinite Queue and the Loss Probability in a Finite Queue," Tech. Rep., Purdue University, West Lafayette, IN, 1999.

[7] A. P. Zwart, "A fluid queue with a finite buffer and subexponetial input," *Advances in Applied Probability*, March 2000, to appear.

[8] J. Choe and N. B. Shroff, "A Central Limit Theorem Based Approach for Analyzing Queue Behavior in High-Speed Networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 659–671, Oct. 1998.

[9] I. Norros, "On the Use of Fractal Brownian Motion in the Theory of Connectionless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 953–962, Aug. 1995.

[10] H. S. Kim and N. B. Shroff, "Loss probability calculations and Asymptotic Analysis for finite buffer multiplexers," Tech. Rep., Purdue University, West Lafayette, IN, 2000.

[11] I. Sidhu and S. Jordan, "Multiplexing gains in bit stream multiplexors," *IEEE/ACM Transactions on Networking*, vol. 3, no. 6, pp. 785–797, Dec. 1995.

[12] J. Choe and N. B. Shroff, "Use of Supremum Distribution of Gaussian Processes in Queueing Analysis with Long-Range Dependence and Self-Similarity," *Stochastic Models*, vol. 16, no. 2, Feb. 2000.

[13] N. G. Duffield and Neil O'Connell, "Large deviations and overflow probabilities for the general single server queue, with application," *Proc. Cambridge Philos. Soc.*, vol. 118, pp. 363–374, 1995.

[14] A. Simonian and J. Guilbert, "Large deviations approximation for fluid queues fed by a large number of on-off sources," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1017–1027, Aug. 1994.

[15] E. W. Knightly and N. B. Shroff, "Admission Control for Statistical QoS: Theory and Practice," *IEEE Network*, vol. 13, no. 2, pp. 20–29, March 1999.

Load anomaly detection for Voice over IP traffic

M. Mandjes, I. Saniee, A.L. Stolyar
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
`{michel,iis,stolyar}@research.bell-labs.com`

Abstract

We consider the problem of traffic anomaly detection in IP networks. We derive general formulae for the variance of the cumulative traffic over a fixed time interval. The main focus of this paper is anomaly detection in voice over IP networks. In this case, the analytical form of the variance formula is very simple. To detect such load anomalies, we show it is sufficient to consider cumulative traffic over relatively long intervals such as 5 minutes. We propose simple anomaly detection tests including detection of over/underload. This approach substantially extends the current common practice in IP network management where only the first order statistics and fixed thresholds are used to identify abnormal behavior. We conclude with the application of the scheme to field data from an operational network.

1 Introduction

IP networks carrying voice traffic are beginning to emerge due to the cost efficiency of IP platforms and their extensibility to other applications and media. However, before this transition becomes widespread, key problems in network management and operation need to be addressed. In this paper we focus on load characterization, overload prediction and more generally *load anomaly detection* in segments of IP networks that carry (almost) exclusively voice traffic, e.g., an egress port of an IP router or switch connected to the trunking voice gateway. As it turns out, many other segments of an IP infrastructure carrying large amounts of voice traffic may be dedicated to carry mostly voice traffic. This emerges from the architectures of many IP-based networks carrying voice.

The need for detection of load anomalies arises, for example, when an atypical load change (increase or decrease) is experienced by a portion of the network. In *focused overload* a large number of callers try to reach the same destination phone number(s) and the network admits too many calls. More generally, overload or indeed *underload* occurs when a segment of the network fails and the traffic either overflows into a normal segment of the network or migrates away from it. In any one of these scenarios, it is desirable to *detect* the ongoing overload or underload as fast as possible, thus helping the network operator to take remedial action or to invoke programmed response. This goal has, unfortunately, turned out to be elusive for the IP networks for a variety of reasons and typically, the network operator hears about performance problems from customers before the large amounts of information available to it has been adequately analyzed. We show that at least in the VoIP segments of an IP network, this need not be the case. Furthermore, it is possible to detect and predict atylical behavior via standard 5-minute traffic measurements.

In an IP-based network, the traffic information available is the cumulative amounts of traffic ("byte counts") over 5 minute time intervals (the *ifInOctets* and *ifOutOctets* MIBs in SNMP). For data traffic five minutes is indeed a long time. Can such "crude" information as 5 minute byte counts be efficiently used for load anomaly detection? We show in this paper that for the VoIP traffic the answer is "yes", under a mild set of assumptions on the coding rate(s) of the packetized voice traffic and the mean call duration.

Our approach is based the analysis of the variance of the byte counts (i.e., "second order" statistics.) Let $A(t_1, t_2)$ denote the cumulative amount of traffic sent on a link in the time interval (t_1, t_2). Expressions for the mean and variance of $A(t_1, t_2)$ can be derived for a very general model of IP traffic, with data sessions arriving as a Poisson process. (We do that in Section 2.1). In the special case of a link with VoIP-only traffic, which is the focus of this paper, the mean and variance of $A(t_1, t_2)$ have very simple closed form expressions, that are described in the Section 2.2. As it turns out, simple tests based on these formulae allow us to discriminate between VoIP and non-VoIP traffic, and detect anomalies in the former.

In summary, the contribution of the paper is the following:

- We argue that second order statistics are useful in traffic anomaly detection and obtain expressions for the variance of general traffic.
- We show that, in the case of VoIP traffic, cumulative byte counts over relatively long intervals can indeed be used for anomaly detection.
- We develop a set of detection procedures, for different types of anomalies.
- We assess the efficacy of the procedures with real traffic traces.

We organized this paper as follows. In Section 2 we, first, derive a formula for the variance of an interval measurment for a genaral IP traffic, and then present explicit formulae for the VoIP traffic with Pareto and exponential distributions of call duration. In Section 3 we apply the variance formulae to devise traffic anomaly detection procedures for VoIP. Section 4 is devoted to numerical validation and illustration of our detection procedures.

2 The variance of an interval measurement

2.1 A general expression

Suppose *data sessions* (which in particular may be voice calls) arrive in time according to a (possibly non-homogeneous) Poisson point process on the real axis $R = (-\infty, \infty)$. (See [6] for the precise definition of a Poisson point process.) Let Λ denote the *intensity measure* of this process, defined on Borel subsets of R. We assume that Λ is finite on bounded sets. This in particular means that the (random) number of arrivals (points) within any finite interval $(t_1, t_2]$ has Poisson distribution with mean $\Lambda\{(t_1, t_2]\}$. (When Λ is absolutely continuous with respect to Lebesgue measure, with the constant density $\lambda = d\Lambda/dt$, then this a gomogeneous process with intensity λ.)

Assume that the (random) amount of traffic generated by a session arrived at time t is described by a random non-decreasing non-negative right-continuous function

$$G_t(x),\ x \geq 0\ ,$$

where $G_t(x)$ is the amount of traffic generated in the closed interval $[t, t+x]$. (Let $G_t(0-) = 0$ by convention.) We assume that the distribution of G_t depends on t as a parameter, and that the functions G_t corresponding to different sessions are independent (even if arrival times of some of them coinside).

Remark. Formally, the process we described is a *marked* Poisson point process [6], with the random functions G being marks of the points. This means that the dependence of the distribution of G_t on t must be such that the marked point process can be well defined as a (measurable) random element. This condition is not restrictive for any conceivable application.

Let us denote for $0 \leq x_1 \leq x_2 < \infty$,

$$f_t^{(1)}(x_1, x_2) \doteq \mathbb{E}\left[G_t(x_2) - G_t(x_1-)\right]\ ,$$

$$f_t^{(2)}(x_1, x_2) \doteq \mathbb{E}\left[G_t(x_2) - G_t(x_1-)\right]^2\ .$$

Now, let $A(t_1, t_2)$ denote the total amount of traffic generated by all sessions in the (closed) interval $[t_1, t_2]$.

Theorem. *For the mean total amount traffic we have*

$$\mathbb{E}A(t_1, t_2) = \int_{-\infty}^{t_2} f_t^{(1)}((t_1 - t)^+, t_2 - t)\Lambda(dt)\ , \tag{2.1}$$

where $x^+ \doteq \max\{x, 0\}$.

If $\mathbb{E}A(t_1, t_2) < \infty$, *then*

$$\operatorname{Var} A(t_1, t_2) = \int_{-\infty}^{t_2} f_t^{(2)}((t_1 - t)^+, t_2 - t)\Lambda(dt)\ . \tag{2.2}$$

Proof sketch. Suppose $X_i, i = 1, 2, \ldots$, is a sequence of i.i.d. random variables, and N is random variable with Poisson distribution, independent of the X_i's. It is well known that

$$\mathbb{E}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}N \cdot \mathbb{E}X_1 \tag{2.3}$$

and

$$\operatorname{Var}\left(\sum_{i=1}^{N} X_i\right) = \mathbb{E}N \cdot \mathbb{E}X_1^2 \tag{2.4}$$

It follows directly from the definition of a marked Poisson point process, that in our case the total contributions into $A(t_1, t_2)$ of the sessions originating in non-overlapping intervals $(a, b]$ and $(c, d]$ are independent. Therefore, it is sufficient to prove the equations (2.1) and (2.2) for the case when the intensity measure is concentrated on a finite interval $(a, b]$.

In this case, our marked point process can be constructed as follows. A Poisson random variable N with mean $\Lambda\{(a, b]\}$ is defined. Also, a sequence of i.i.d. random "extended" marks is defined, where each mark is a random function $((G_t(x), x \geq 0), t \in (a, b])$ of *two* variables x and t. A realization of our process is constructed by first taking a realization of N, then placing N points into the interval $(a, b]$ independently according to the distribution $\Lambda(dt)/\Lambda\{(a, b]\}$, and finally "attaching" first N extended marks (from the i.i.d. sequence) to the points. The mark $(G_t(x), x \geq 0)$ of a point located at t is simply the projection of its extended mark at time t.

Then, formulas (2.1) and (2.2) follow directly from (2.3) and (2.4), respectively.

2.2 VoIP formulas

Suppose now that the data sessions are voice calls with i.i.d. durations, having the distribution function $H(x), x \geq 0$, with the density $h(x), x \geq 0$, and finite mean μ^{-1}. Assume that calls arrive according to a homogeneous Poisson process with intensity $\lambda > 0$, and that each call in progress generates data at the constant rate 1.

In this case the distribution of the amount of traffic $A(t_1, t_2)$ depends only on $t_2 - t_1$, so we will write $A(t) \doteq A(0, t)$, and obviously $\mathbb{E}\, A(t) = (\lambda/\mu)t$. Although the general formula (2.2) can be applied directly to obtain $\mathbb{V}\text{ar}\, A(t)$, in our special case the derivation can be somewhat simplified by using the well known fact that (in a $M/GI/\infty$ queueing system) the random number N_1 of calls in progress at time 0 has Poisson distribution with mean λ/μ, and the residual call durations are i.i.d. with the density $h^*(x) = \mu(1 - H(x))$ and the distribution function $H^*(x) = \int_0^x h^*(y)dy$. Also, the number of calls which originate in $[0, t]$ is Poisson with mean $N_2 = \lambda t$.

If we denote

$A_1(t) :=$ traffic generated in $[0, t]$ by flows that were already present at time 0,

$A_2(t) :=$ traffic generated in $[0, t]$ by flows that entered during the interval,

then

$$\mathbb{V}\text{ar}\, A(t) = \mathbb{V}\text{ar}\, A_1(t) + \mathbb{V}\text{ar}\, A_2(t)\,,$$

where (applying (2.4))

$$\mathbb{V}\text{ar}\, A_1(t) = (\lambda/\mu) \int_0^t u^2 h^*(u)du + t^2(1 - H^*(t)),$$

and

$$\begin{aligned} \frac{\mathbb{V}\text{ar}\, A_2(t)}{\lambda t} &= \int_0^t \int_0^{t-u} \frac{s^2}{t} h(s)dsdu + \int_0^t \int_{t-u}^{\infty} \frac{(t-u)^2}{t} h(s)dsdu \\ &= \int_0^t \int_0^{t-u} \frac{s^2}{t} h(s)dsdu + \int_0^t \frac{(t-u)^2}{t} (1 - H(s))du. \end{aligned}$$

Consider two special cases.

- In the first, the call duration is exponential, i.e., $h(x) = \mu \exp(-\mu x)$. Straightforward calculations yield

$$\mathbb{V}\text{ar}\, A(t) = \frac{2\lambda t}{\mu^2} - \frac{2\lambda}{\mu^3}(1 - e^{-\mu t}). \tag{2.5}$$

If we denote by $\bar{A}(t) \doteq A(t)/t$ the (random) average number of calls in an interval of length t, by $a = \lambda/\mu$ the mean number of calls, and by $T = \mu t$ the measurement interval length normalized by the mean call duration, then we can rewrite (2.5) in a "normalized" form:

$$\mathrm{Var}\,\bar{A}(t) = \sigma_T^2(a) \doteq 2a\frac{e^{-T} - 1 + T}{T^2}. \tag{2.6}$$

This is the formula found by Riordan [5]. (The extension to the case when there is only a finite number of "trunks" available for calls was obtained by Beneš [2].)

- Now assume that a call duration has a Pareto distribution. More precisely, for $s \geq 0$,

$$1 - H(s) = \frac{1}{(s+1)^{\alpha}}, \quad 1 - H^*(s) = \frac{1}{(s+1)^{\alpha-1}}.$$

With similar calculations,

$$\mathrm{Var}\,A(t) = \frac{2\lambda}{(1-\alpha)(2-\alpha)(3-\alpha)}\left(1 - \frac{1}{(t+1)^{\alpha-3}}\right) + \frac{2\lambda t}{(1-\alpha)(2-\alpha)}.$$

Related calculations for Gaussian sources were done in [1].

3 Alarm types

Consider a VoIP link, and assume that call durations do have exponential distribution. The Riordan formula turns out to be useful in helping detect and predict high load and abnormal traffic behavior. To this end, a set of tests are constructed using the Riordan formula as the underlying model for variation of the load data measured every 5 minutes. Typically, three conditions need to be checked when one applies a model to data. First, the applicability of the model to measurements needs to be tested. This is to check that the data meets the assumptions made in the model. Second, a test needs to determine if measurements indicate normal or abnormal load behavior, given that the data pass the first test. Third, a final test predicts if overload will be in effect. These tests, or *alarms*, are detailed below.

I. Traffic Type Alarm. This alarm is issued when the *medium-term* empirical mean and variance of the measurements, or byte-counts $(M_i)_{i\in\mathbb{N}}$, taken over a fairly long sequence (say a few hours) of 5-minute intervals differs substantially from the theoretical mean and variance predicted by the Riordan formula. Such an alarm indicates *consistency* or 'conformance' of the measurement statistics with those of VoIP traffic.

To be more precise, we propose the following procedure. Consider n subsequent byte counts (over periods of length t=5 min) $M_1, \ldots, M_n$. Let $\bar{M}_n$ be their average. Compute the (normalized) empirical variance

$$\hat{\sigma^2} := \frac{1}{t^2}\frac{1}{\gamma^2}\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M}_n)^2,$$

where γ is the data rate generated by one call in progress. Now compare $\hat{\sigma^2}$ with its theoretical value $\sigma_T^2(a)$, according to the Riordan formula (2.6).

Possible reasons for non-consistency are:

- A significant fraction of the traffic is *not* VoIP. Evidently, if user data sessions do not generate a constant rate traffic (as voice calls do), Riordan formula will fail.
- Even if traffic is VoIP, Riordan formula does not necessarily hold. In the first place, it assumes that the call holding time has exponential distribution. But even if the call durations are indeed exponential, the above test might fail. The reason for this fact is that, as seen from (2.6), estimation of the variance requires that the *mean* holding time is known. A misspecification of μ obviously leads to a wrong estimate.
- The test might fail due to the inherent non-stationarity of data. This problem could be circumvented by removing the trends in the data. This can be done in a straightforward way.
- Evidently, non-consistency can be caused by continual misleading or false measurements.

In case of such an inconsistency, the alarms of types II and III described below, which *assume* measurements are VoIP, should be ignored because the current test indicates lack of agreement between observed variability of measurements and the model, thus invalidating the applicability of the model to the measurements at hand.

II. Fast Load Change Alarm. This alarm is issued when the current byte count M, or the derived number of 'calls', is outside *either above or below*, the (say) 95% confidence interval predicted by Riordan formula with the mean load a set to the *short-term* empirical mean load taken over 2-4 last measurements. As noted before, this procedure has to be executed only if no traffic type alarm is issued, in other words, there is no suspicion that the traffic is not VoIP.

The alarm may be caused by a number of events. A first possible cause is a link failure somewhere in the network. This failure triggers rerouting of calls, thus leading to an effective load increase or decrease. Another reason may be a sudden traffic increase due to an 'external' event, i.e., an event unrelated to the network. Also a measurement failure might lead to an alarm.

Remark: Consistent violations. If the percentage of measurements that fall outside the adaptively predicted 95% confidence band is substantially more than 5%, then this is ground to suspect applicability of the model to the measurement. This is the same situation as a Type I alarm, and indeed this approach provides an alternative method to set an alarm of Type I.

III. Link Overload Alarm. This alarm indicates that the link load is too high. It is issued when the current byte count (or the derived number of calls) exceeds a calculated threshold value, $M > M^*$, where M^* is a predefined threshold, and the first alarm is not set. The value of M^* is static, it is a fixed function of the link speed C and we show below how it is computed. Typically, this alarm requires some immediate action as link overload would result in packet losses and consequent voice quality deterioration. We note that static threshold alarms, such as this, are the most commonly used in network management. The main difference being that the threshold in this case is derived from the model and its estimated parameters and not empirically set. In other words, the model gives a threshold that can be used as a guide for setting an empirical overload threshold.

Suppose that the link speed C is large enough, say at least the speed of a T3 port, which corresponds to about 500 calls at 64 kbit/s. This guarantees that at the loads of the order of the link speed, the distributions of M_i and the instantaneous data rate are approximately normal. Suppose, we are given a constant $\beta > 0$ which is the maximum acceptable probability of the *instantaneous* data rate exceeding the link speed C. (Typically, β is 0.01 or 0.05.) Then the maximal acceptable link load a^* (i.e., the mean number of calls in progress) is computed from

$$a^* + b\sqrt{a^*} = C/\gamma \; .$$

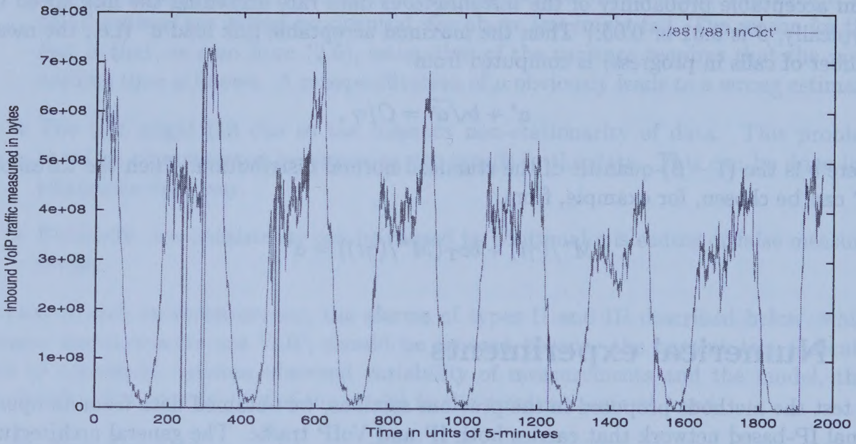
where b is the $(1-\beta)$-quantile of the standard normal distribution. Then the threshold M^* can be chosen, for example, from

$$M^*/(\gamma t) + b\sigma_T(M^*/(\gamma t)) = a^* \; .$$

4 Numerical experiments

To test the methods proposed in the previous sections, we obtained data from an operational IP-based network that carries both IP and VoIP traffic. The general architecture of the network indicates that many segments carry primarily VoIP traffic. The data contain 5-minute input/output byte measurements per interface (*ifInOctets* and *ifOutOctets* SNMP MIBs) collected over a period of one week. The resulting byte counts from the two typical interfaces, one for each of IP and VoIP loads, are shown in Figure 1.

We apply the tests proposed in the previous Section to the available traces. First, for the applicability of the Riordan method to these data (Type I Test), 30-minute sample variances are normalized by removing any trends and rescaled to match a 1 erlang load. Figure 2 shows the plot of the normalized sample variances of the VoIP process versus the expected variance from Riordan's formula. As it can be seen from this graph, there are 25 occasions within the week when the variance exceeds the theoretical limit and the average time to exceed the Riordan threshold is ~400 minutes. Further, these intervals are spread to the right with a heavy tail. In contrast, the IP traffic exceeds the limit on ~90 occasions with the average time before exceeding the threshold being ~90 minutes, with intervals closelt clustered around the mean. We conclude that the model applies to VoIP traffic, but not IP traffic.



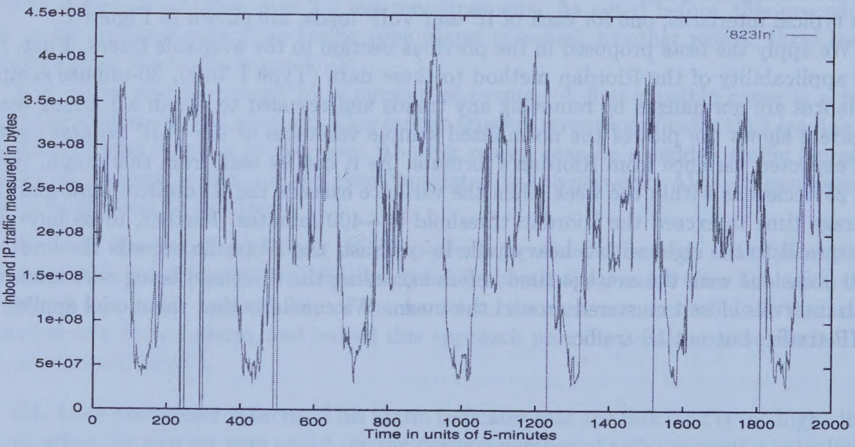


Figure 1: Profile of VoIP traffic (top) and IP traffic (bottom) from an operational network, as measured by the total incremental bytes every 5 minutes for a week.

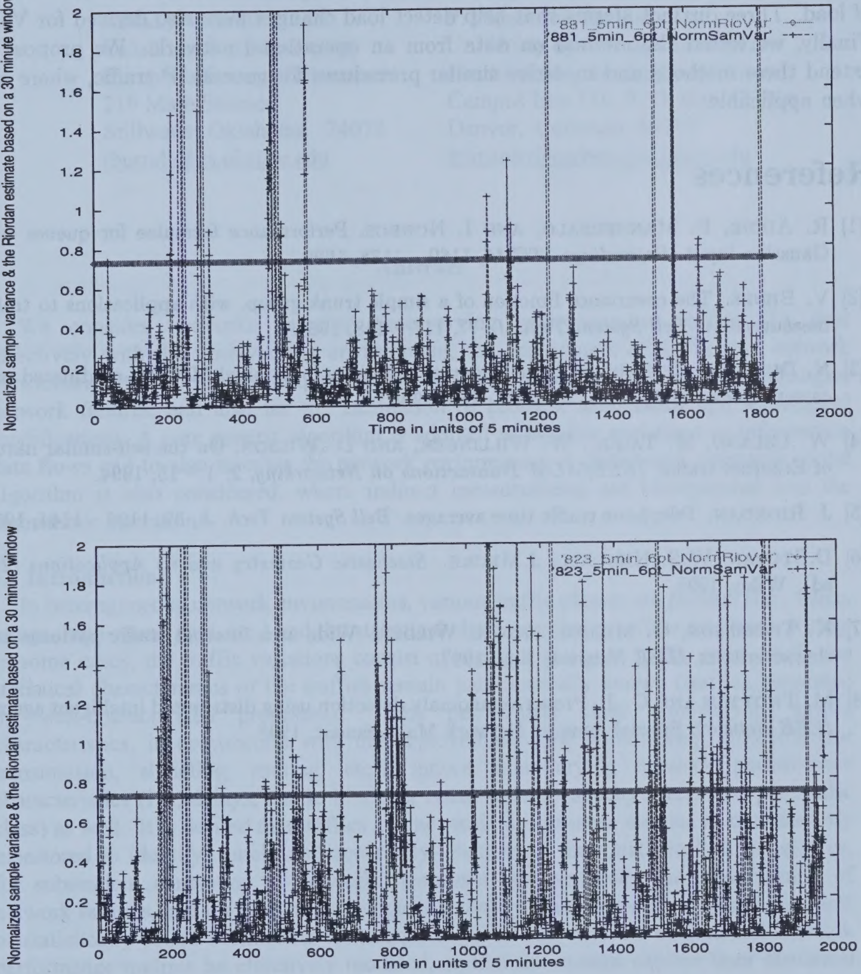


Figure 2: Comparison of the sample variance with a window size 6 to the theoretical (Riordan) variance for the normalized VoIP traffic (top) and IP traffic (bottom).

5 Conclusions

We identified procedures for detection of load anomalies using measurements over fixed time intervals in VoIP network segments. This was done by estimating the mean and variance of load and derivation of general analytical expressions for the second moment of load. Three distinct alarms that help detect load changes were also derived for VoIP. Finally, we tested the method on data from an operational network. We propose to extend these methods and to derive similar procedures for general IP traffic, where and when applicable.

References

[1] R. Addie, P. Mannersalo, and I. Norros. Performance formulae for queues with Gaussian input. *Proceedings ITC 16*, 1169 – 1178, 1999.

[2] V. Beneš. The covariance function of a simple trunk group, with applications to traffic measurement. *Bell System Tech. J.*, ??, 117 – 148, 1961.

[3] N. Duffield. Queueing at large resources driven by long-tailed M/G/∞-modulated processes. *Queueing Systems,* 28: 245 – 266, 1998.

[4] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking*, 2: 1 – 15, 1994.

[5] J. Riordan. Telephone traffic time averages. *Bell System Tech. J.*, 39, 1129 – 1144, 1951.

[6] D.Stoyan, W.S. Kendall, J.Mecke. *Stochastic Geometry and its Applications* (2nd ed), Wiley, 1995.

[7] K. Thompson, G. Miller, and R. Wilder. Wide area internet traffic patterns and characteristics *IEEE Network*, Nov. 1997.

[8] M. Thottan and C. Ji. Proactive anomaly detection using distributed intelligent agents, *IEEE Network*, Special issue on Network Management, 1998

A GENERIC ALGORITHM FOR MONITORING VARIOUS STOCHASTIC NETWORK CHARACTERISTICS

Anthony Burrell
Oklahoma State University
Computer Science Department
219 Math Science
Stillwater, Oklahoma 74078
tburrell@cs.okstate.edu

P. Papantoni-Kazakos
University of Colorado at Denver
Electrical Engineering Department
Campus Box 110, P. O. Box 173364
Denver, Colorado 80217
tpapanto@carbon.cudenver.edu

Abstract

We consider networks carrying time-varying heterogeneous traffics. To deal effectively with the time-varying environment, the deployment of traffic and network performance monitoring techniques is necessary for the identification of traffic changes, network failures, and also for the facilitation of protocol adaptations and topological modifications. A core generic algorithm is adopted to monitor variations in information data flows and to also monitor the network performance. A distributed extension of the algorithm is also considered, where indirect measurements are incorporated into the algorithmic operations.

1. Introduction

In heterogeneous network environments, various traffic classes are present (i.e., voice, image, high-speed data, etc.) and the statistical characteristics per class are time-varying. In some cases, the traffic variations consist of changes solely in traffic rates, while the statistical characteristics of the traffics remain parametrically known (traffics generated by well-characterized processes). The per traffic class time-varying traffic characteristics, in conjunction with the deployed network multiplexing protocols (for transmission, signaling, routing, etc.), induce time-varying network performance characteristics (i.e., delays, traffic blocking rates, wasted capacity rates, etc., per traffic class) as well. It is desired that traffics and network performance metrics be continuously monitored to identify, timely and accurately, changes in their statistical characteristics, for subsequent adaptation of network operations and protocols, for identification of network failures, and for possible topological reconfigurations as well. The development of statistically sound monitoring techniques requires that first the traffics and network performance metrics be effectively modeled, so that the models capture their statistical variations. We model them as sets of distinct and well-known stochastic processes (e.g., Poisson with known rates, geometric with known parameters, etc.). For example, for the k-th traffic class we consider the general scenario where the various possible arrival processes that characterize the traffic may be represented by the set $\{\mu_i^{(k)}\}_{1\le i\le m}$ where $\mu_i^{(k)}$ depicts a well-known process, such as a Poisson process with known fixed rate, and where all processes in the set are distinct (e.g., all Poisson with different and distinct rates); it is assumed that the set $\{\mu_i^{(k)}\}_{1\le i\le m}$ has been obtained via thorough traffic characterization methods. The objective of monitoring the k-th traffic class is then to identify, timely and accurately, shifts from some process $\mu_j^{(k)}$ in the set to any other process in the same set. In section 2, we present a core monitoring algorithm that operates on models represented by sets of distinct stochastic processes. In section 3, we present a distributed extension of the algorithm in section 2. In reference [1], [2] and [3], pertinent network applications can be found, while in reference [4] some theoretical results are included.

2. Core Monitoring Algorithm

We first represent an initial restricted form of the algorithm and then a reinitialization extension. The latter actually represents the core monitoring algorithm that we propose. Both the initial restricted form and the reinitialization extension operate on observed data sequences $x_1, x_2, \ldots, x_n, \ldots$ that are generated by the acting processes (e.g., in the case of processes representing traffics, the x_i's may be numbers of arrivals within consecutive fixed-length time intervals). Below, we use the abbreviated form x_1^n, for the sequence $x_1, \ldots, x_n$.

Initial Restricted Algorithm

The restricted algorithm addresses the following problem.

Let the process which initially generates the data sequence be the process μ_0. Let it be possible that a shift to any one of m-1 independent processes μ_i ; i=1, ..., m-1 may occur at any point in time, where if a $\mu_0 \rightarrow \mu_i$ shift occurs, then the process μ_i remains active thereafter. The objective is to detect the occurrence of a $\mu_0 \rightarrow \mu_i$ shift as accurately and as timely as possible, including the detection of the process μ_i which μ_0 changed to. Let us denote by f_i ; i=0, 1, ...,
m-1 density or probability functions induced by the processes μ_i ; i=0, 1, ..., m-1. Then, for the present problem, we propose the following algorithm.

Algorithm

(a) Select a threshold $\delta_0 > 0$.

(b) Have m-1 parallel algorithms operating. The ith algorithm; i =1, ..., m-1 is monitoring a $\mu_0 \rightarrow \mu_i$ shift. $T_n^{0i}(x_1^n)$ denotes the operating value of the ith algorithm at time n, given the observation sequence x_1^n. The operating value $T_n^{0i}(x_1^n)$ is updated as follows.

$$T_0^{0i} \equiv 0$$

$$T_n^{0i}(x_1^n) = \max\left(0,\ T_{n-1}^{0i}(x_1^{n-1}) + \log \frac{f_i(x_n|x_1^{n-1})}{f_0(x_n|x_1^{n-1})}\right) \qquad (1)$$

(c) The algorithmic system stops the first time n when either one of the m-1 parallel algorithms crosses the common threshold δ_0. If the ith algorithm is the one that first crosses the threshold, then it is declared that a $\mu_0 \rightarrow \mu_i$ shift has occurred. □

The algorithm described above is clearly sequential and thus computationally simple and efficient. Its fundamental performance characteristics are given by the Theorem and the Corollary given below, where first some quantities are defined.

Let N_{0i} denote the extended stopping variable induced by the ith algorithm in the system; that is, $N_{0i} \triangleq inf\left\{n: T_n^{0i}(x_1^n) \geq \delta_0\right\}$. Let us define the following quantities.

$$L_n^{ij}(x_1^n) \triangleq n^{-1} \log \frac{f_j(x_1^n)}{f_i(x_1^n)} \qquad ; i, j = 0,\ 1,\ ..,\ m-1$$

$$I_{ij} \triangleq \lim_{n\to\infty} L_n^{ij}(x_1^n) \qquad ; i, j = 0,\ 1,\ ..,\ m-1$$

$$p_n^{ij}(v) \triangleq P\left(L_n^{ij}(x_1^n) < v \,|\, \mu_j\right) \qquad ; i, j = 0,\ 1,\ ..,\ m-1$$

Consider then, the following conditions:

(A) $I_{ij}; i, j = 0,1,..,m-1$ exist and

$I_{ij} = E\left\{I_{ij} | \mu_j\right\}$, a.s. (P_{μ_j})

(B) For $i, j = 0,\ 1,\ ..,\ m-1$ and for $v \in (0, I_{ij})$,

$\lim_{n\to\infty} n\, p_n^{ij}(v) = 0$ and $\sum_{n\geq 1} p_n^{ij}(v) < \infty$

Conditions (A) state the existence of the generalized Kullback-Leibler information numbers $\left\{I_{ij}\right\}$ while conditions (B) ensure that the convergence of the large-deviation probabilities $\left\{p_n^{ij}(v)\right\}$ is sufficiently fast. The following theorem and corollary can now be expressed.

Theorem 1

Let the processes $\{\mu_j ; j = 0,\ 1,\ ..,\ m-1\}$ be stationary, ergodic, mutually independent, and satisfying conditions (A) and (B). Then,

$$\text{As } \delta_0 \rightarrow \infty,\ \underset{\substack{j=0,\ 1,\ ..,\ m-1 \\ i=1,\ ...,\ m-1}}{E\{N_{0i} | \mu_j\}}: \begin{cases} \sim \left[I_{0j} - I_{ij}\right]^{-1} \log \delta_0 & ;\ \text{if } I_{0j} \geq I_{ij} \\ \geq 2^{-1} \delta_0 & ;\ \text{if } I_{0j} < I_{ij} \end{cases}$$

Corollary

Given that the process μ_j ; $j = 1,\ ..,\ m-1$ is acting throughout, the algorithmic system will asymptotically detect the $\mu_0 \rightarrow \mu_j$ shift correctly, in the expected stopping time sense. That is:

$$E\left\{N_{0j} | \mu_j\right\} \sim I_{0j}^{-1} \log \delta_0 : \begin{cases} < E\left\{N_{0i} | \mu_j\right\} \sim \left[I_{0j} - I_{ij}\right]^{-1} \log \delta_0 & ;\ \forall i \neq j : I_{0j} > I_{ij} \\ < E\left\{N_{0i} | \mu_j\right\} \geq 2^{-1} \delta_0 & ;\ \forall i \neq j : I_{0j} < I_{ij} \end{cases}$$

Given that the process μ_j ; j =1, .., m-1 is acting throughout, the asymptotic expected stopping time of the algorithmic system is: $E\{N_{0j} \mid \mu_j\} \sim I_{0j}^{-1} \log \delta_0$.

Remarks: We note that the performance of the algorithm depends on the values of the Kullback-Leibler numbers $\{I_{ij}\}$ which represent a measure of "closeness" between the different processes in the model. The "closer" to each other the processes are in that sense, the worst is the performance of the algorithmic system.

The algorithm presented in this Section can be extended to reinitialize each time a decision is being made, to detect shifts from any one of the processes $\{\mu_j \,; j=1, .., m\text{-}1\}$ to one of the remaining ones (see reference [4]).

3. A Distributed Extension

Let us now consider the case where, in addition to the local data, the monitoring algorithm (MA) also utilizes data from other network nodes, in an indirect fashion. Specifically, let the algorithm also utilize decisions induced by the monitoring algorithms (MAs) employed at neighboring nodes. Let the number of the neighboring nodes be N, indexed from 1 to N, and let $u_{i,\,1\text{-}j}$; j = 0, 1, i =1, ..., N, be defined as follows:

$$u_{i,1-j} = \begin{cases} 1; \text{if the ith node decides that} \\ \text{the process } \mu_{1-j} \text{ is active} \\ 0; \text{if the ith node decides that} \\ \text{the process } \mu_j \text{ is active} \end{cases} \tag{2}$$

We then propose a Distributed Monitoring Algorithm (DMA) whose only difference from the centralized MA lies in the updating step. Indeed, we propose that the log ratio in (1) be substituted by the updating step,

$$\sum_{i=1}^{N} w_i u_{i,j} + \log \frac{f_j\left(x_n \mid x_1^{n-1}\right)}{f_0\left(x_n \mid x_1^{n-1}\right)} \tag{3}$$

where $\{w_i\}$ are constants whose objective is to weigh the contribution of the various MAs according to their respective performance characteristics. Let, for example, the MAs monitor link traffics and let then α_i and β_i denote respectively the traffic rejection rate and the wasted capacity rate induced by the MA employed by the ith neighboring node. Then, drawing from some parallelisms with the models considered in [5], we propose the following form of the set $\{w_i\}$:

$$w_i = \log \frac{(1-\beta_i)(1-\alpha_i)}{\beta_i \alpha_i} ; i = 1, \ldots, N \tag{4}$$

We note that the weights $\{w_i\}$ in (4) are positive iff the sum of the traffic rejection and the wasted capacity rates is less than one. We also note from (3) that the decisions of a neighboring MA contribute then positively to the acceleration of a $\mu_{1-j} \to \mu_j$ shift decision (by increasing the size of the updating step), when these decisions point to the μ_j process as being active.

Let $N_{1-j,j}$; j=0,1 denote the extended stopping variable induced by the DMA algorithm with updating step number as in (3). Let also $I_{1-j,j}$ denote the Kullback-Leibler number of the process μ_j with respect to the process μ_{1-j}. Let $N^{(i)}_{1-j,j}$; $j = 0,1$ denote the extended stopping variable induced by the MA deployed by the i^{th} neighboring node and let us define:

$$\rho_{i,k/j} \equiv \frac{E\{N^{(i)}_{k,1-k} / \mu_j\}}{E\{N^{(i)}_{k,1-k} / \mu_j\} + E\{N^{(i)}_{1-k,k} / \mu_j\}} \; ; \; {}^{k=0,1}_{j=0,1} \qquad (5)$$

Then, the following theorem can be expressed.

Theorem

Let the processes $\{\mu_j \; ; \; j = 0, 1\}$ be stationary, ergodic, mutually independent and satisfying the general mixing conditions (A) and (B) in reference [4]. Then, the asymptotic performance of the DMA is as follows:

$$E\{N_{1-j,j} | \mu_j\} \approx \frac{\log \eta_{1-j}}{\sum_{i=1}^{N} w_i \, \rho_{i,j/j} + I_{1-j,j}} \quad , \text{as} \, \eta_{1-j} \to \infty \qquad (6)$$

$$E\{N_{1-j,j} | \mu_{1-j}\} : \begin{cases} \approx \left[\sum_{i=1}^{N} w_i \, \rho_{i,j/1-j} - I_{j,1-j} \right]^{-1} \log \eta_{1-j} ; \text{if} \sum_{i=1}^{N} w_i \, \rho_{i,j/1-j} > I_{j,1-j} \\ \geq 2^{-1} \eta_{1-j} \; ; \text{if} \sum_{i=1}^{N} w_i \, \rho_{i,j/1-j} < I_{j,1-j} \end{cases} \quad , \text{as} \, \eta_{1-j} \to \infty \qquad (7)$$

At the same time, the performance of the MA in (1) is as shown below.

$$E\left\{N_{1-j,j} \middle| \mu_j\right\} \approx \frac{\log \eta_{1-j}}{I_{1-j,j}}, \text{ as } \eta_{1-j} \to \infty \qquad (8)$$

$$E\left\{N_{1-j,j} \middle| \mu_{1-j}\right\} \geq 2^{-1} \eta_{1-j}, \text{ as } \eta_{1-j} \to \infty \qquad (9)$$

□

The results in the Theorem exhibit clearly the interactive relationship between the DMA and the MAs deployed by the neighboring nodes.

4. Conclusions

The core algorithm presented and its distributed extension are generic in the sense that they give rise to a variety of monitoring protocols that are applicable at various levels in the logical structure of a network. The algorithms possess superior performance characteristics. When deployed they may assist in the implementation of superior traffic and network management systems.

5 References

[1] A.T. Burrell, “Traffic Management in Broadband Integrated Services Digital Networks”, Ph.D. Dissertation, Univ. of Virginia, August 1994.

[2] A.T. Burrell, D. Makrakis, and P. Papantoni-Kazakos, “Traffic Monitoring for Capacity Allocation of Multimedia Traffic in Broadband Networks”, *Telecommunications Systems Journal*, 1998, volume 9, pp. 173-206.

[3] A.T. Burrell, H.P. Stern, and P. Papantoni-Kazakos, “Dynamic Capacity Allocation and Hybrid Multiplexing Techniques for ATM Wireless”, *Wireless Networks Journal*, 1998, Vol. 3, pp. 307-316.

[4] A.T. Burrell and P. Papantoni-Kazakos, “Extended Sequential Algorithms for Detecting Changes in Acting Stochastic Processes”, *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 28, No. 5, Sept. 1998, pp. 703-710.

[5] D.A. Pados, P. Papantoni-Kazakos, D. Kazakos and A. Kogiantis, “On-Line Threshold Learning for Neyman-Pearson Distributed Detection”, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 24, No. 10, pp. 1519-1531, Oct. 1994.

Limits of BER Guarantees in Transparent Optical Networks Using OOK Signalling

Poompat Saengudomlert and Muriel Médard
Massachusetts Institute of Technology

5 October 2000

Abstract

A network consisting of transparent optical nodes (TONs) can provide high speed end-to-end communication paths with very low bit error rates (BERs). However, owing to component crosstalk and other degradations at TONs, the BER of a particular communication path traversing several TONs can be degraded by a few orders of magnitude even in the absence of component failure. Monitoring the quality of service (QoS) of a communication path has typically relied on sporadic BER testing and operation monitoring by the nodes using probe signals. Intermittent BER testing cannot provide continuous monitoring of the network QoS. On the other hand, the use of probe signals is not sensitive enough to detect the BER degradation. Based on a new method of monitoring service degradation at individual nodes proposed in [1, 2], we propose a modification and derive the BER value that this modified method can guarantee in the presence of signal degradation due to coherent crosstalk at TONs.

1 Introduction

Consider a communication path traversing several network nodes. In an electronic network, each intermediate node (between transmitting and receiving nodes) generally detects and regenerates (retransmits) traversing data signals. In an optical network, a high optical link bandwidth allows a transmission rate so high that electronic processing of data units (e.g. IP packets, ATM cells) cannot be done in real time. For example, an ATM cell of 53 bytes must be processed under 42.4 ns if an optical link transmits at 10 Gb/s. Consequently, it is desirable to have certain intermediate nodes be transparent optical nodes (TONs). A TON is a node which does not perform optical-to-electronic or electronic-to-optical conversion. Thus, signals traversing TONs remain in the optical domain. An end-to-end communication path through several TONs is referred to as a light path. A communication service through a light path can be made transparent to the formats (protocols) of transmitted data. However, since TONs do not detect and regenerate data signals, there is no error detection and correction of data signals at intermediate nodes. Data signals experience cumulative effects of degradation at individual TONs along the light path.

Common TONs are wavelength selective switches (WSSs), optical amplifiers, optical multiplexers and demultiplexers. Certain switching TONs exhibit crosstalk effects [3, 4, 5, 6], namely a light signal designated (switched) to one output fiber leaks onto another output fiber. Optical amplifiers are used to compensate for power loss within the network. The most common amplifiers are Erbium doped fiber amplifiers (EDFAs),which are broadband noise sources in the communication path. In addition, EDFAs exhibit gain competition, namely signals on different wavelengths must compete for power resources. An unusually high-powered signal on one wavelength can result in lower amplifier gains for all other wavelengths sharing the EDFA.

A data signal traversing TONs can be degraded as a result of crosstalk and gain competition at TONs. Figure 1 illustrates coherent crosstalk in a light path going through two TONs. Coherent crosstalk occurs when crosstalk signals have the same frequency as data signals and thus directly alter the data signal waveforms. Neglecting signal attenuation and all noise sources, the receiver receives the data signal s_d together with a crosstalk signal s_{c1} introduced at TON 1 by light path c_1 and a crosstalk signal s_{c2} introduced at TON 2 by light path c_2. We concentrate our analysis on coherent crosstalk since non-coherent crosstalk due to gain competition is device specific and is outside the scope of this work.

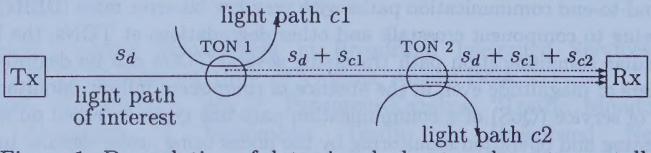


Figure 1: Degradation of data signals due to coherent crosstalk.

Section 2 describes existing supervisory methods that may be used to monitor the QoS of a communication path. We argue that these methods are not sufficient to detect small BER degradations. Section 3 describes a novel BER monitoring system proposed in [1, 2] and provides the setup for the analysis that follows. Section 4 contains the analysis leading to the BER guarantees provided by the system discussed in Section 3. To improve the sensitivity of the system, we proposes a modification using hard limiters at TON inputs. Section 5 provides a summary and directions of future research.

2 Existing Supervisory Methods

This section briefly discusses existing supervisory methods that may be used to detect BER degradation. [7] provide detailed explanation on this topic. Two main approaches exist for monitoring the light path QoS. The first relies on the use of pilot signals, which are known signals inserted to travel along the same paths as data signals but are distinguishable from data signals. As an alternative, we can use an optical time domain reflectometer (OTDR) to detect the echo of pilot signals at the transmitter. The second approach is based on observing the statistical parameters such as optical signal average power, optical signal frequency spectrum, and the BER of a light path.

Coherent crosstalk can degrade the BER without causing degradation to pilot signals. In the case that pilot signal frequencies are close to those of data signals, they are referred to as subcarrier multiplexed (SCM) signals. In some cases, SCM signals are recovered by detecting them from the superposition with data signals. However, SCM signals are generally modulated at a much lower rate than data signals [8, 9]. Therefore, a small percentage of data bits can be degraded without appreciable degradation of SCM signals.

Consider the second supervisory mechanism based on observing statistical parameters. A power detector may be used to detect BER degradation if crosstalk signals change the average power of data signals. If a crosstalk signal at TON 1 is present only 1% of the time, the BER may deteriorate by several orders of magnitude (up to 10^{-2}) while the average received signal power does not change significantly. By the same arguments, an optical spectrum analyzer (OSA) cannot detect BER degradations due to a sporadic crosstalk signal which is present only at a small percentage of time.

The use of a BER tester (BERT) is a standard procedure to detect BER degradation. However, BER testing is only performed sporadically on test patterns. Moreover, the use of a BERT requires a long detection time. For example, a BERT takes several hundreds of seconds to distinguish between a BER less than 10^{-10} and a BER greater than 10^{-8} for a transmission rate of 1 Gb/s.

Short detection time is desirable in several aspects. Since optical transmission rates are extremely high, a large amount of data can be affected before any action takes place. If retransmission is to be performed after the detection of excessive BER, long detection time implies the use of a large amount of buffer storage. As we shall see, our proposed method of detecting BER degradation is based on the detection of signal degradation at TONs rather than the detection of error bits themselves. As a result, our method can operate at roughly 6 orders of magnitude faster than a BERT.

3 Novel Method for Detecting BER Degradation

This section describes the method proposed in [1, 2] for detecting BER degradation in a light path. The detection system is constructed by installing, at each TON along a light path, a service monitoring device which wraps around its TON as shown in figure 3. The device is wavelength selective and compares the signals from input and output taps on a particular WDM channel. The relation between these two signals yields a diagnostic of TON operation. Our analysis will concentrate on a single WDM light path using OOK signalling with coherent detection.

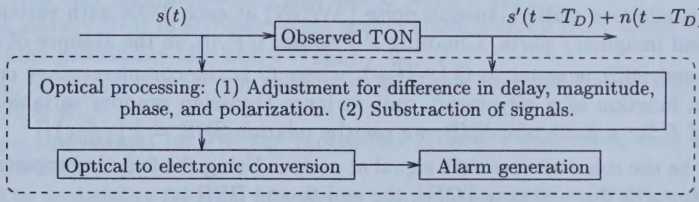

$s(t)$
$s'(t - T_D) + n(t - T_D)$
Observed TON
Optical processing: (1) Adjustment for difference in delay, magnitude, phase, and polarization. (2) Substraction of signals.
Optical to electronic conversion
Alarm generation

Figure 2: A wrapping service monitoring device at a TON.

From figure 3, denote the input signal by $s(t)$. Denote the output signal by $s'(t-T_D)$, where T_D is the TON delay and s' is used instead of s in consideration of the changes in magnitude (tap losses and amplification), phase, and polarization. Denote the sum of signal and noise at TON output by $s'(t - T_D) + n(t)$. In the optical domain, the device takes two tapped signals and compensates for the difference in delay, magnitude, phase and polarization under normal operation. It then performs signal subtraction and optical to electronic conversion using coherent detection. In general, the magnitude, phase, and polarization of data signals do not vary rapidly; and we can assume, in the absence of crosstalk, that the device yields an output proportional to $n(t)$. For notational simplicity, we shall drop the time notation in what follows.

If there is a degradation, denoted by J, introduced at the TON, the device output will be proportional to $n + J$. When the output exceeds a preset threshold, the device generates an alarm notifying excessive signal degradation at the TON. With such a wrapping device installed at each node along the light path, we consider, in each bit transmission period, that a "device alarm" is generated if at least one TON in the light path generates an alarm. In each observation period, we count the total number of device alarms. If this total is higher than a preset threshold, a "BER alarm" (notifying an excessive BER) is generated for that observation period.

The performance criteria of the BER monitoring system are the false positive probability (FP), the probability of generating a BER alarm when the BER is not excessive, and the false negative probability (FN), the probability of not generating a BER alarm when the BER is excessive. Preliminary results on the performance of this BER monitoring system are given in [2]. However, the results in [2] are based on direct detection of signals and the assumption that coherent crosstalk signals are constant throughout an observation period and are equal at all TONs. In this work, we relax these constraints and investigate the performance of this approach for monitoring the BER. In particular, we want to provide a BER guarantee for each observation period in the absence of an BER alarm (with some value of FN) regardless of the nature of signal degradation at TONs. To do so, we derive the worst case scenario with the lowest probability of a BER alarm given a fixed BER. Specifically, if the BER monitoring system can detect the worst case scenario corresponding to the BER of $\tilde{B}$, we can guarantee that the BER is less than $\tilde{B}$ in the absence of a BER alarm.

3.1 Setup for the Analysis

Let s_R and s_I denote the real and imaginary parts of a signal s. Let s^{OFF} and s^{ON} denote the OFF and the ON signals for OOK signalling. We shall assume that the OFF and the ON signals are equally likely. By expressing s in the form (s_R, s_I), we can write s^{OFF} as (0,0) and s^{ON} as $(s_R^{\text{ON}}, s_I^{\text{ON}})$. Let M denote the number of TONs in the light path. Assume that there is additive white Gaussian noise (AWGN) at each TON with variance σ^2 for both real and imaginary parts. Choosing s^{ON} to be $(\sqrt{P}, 0)$, in the absence of crosstalk, the end-to-end BER is equal to $Q\left(\frac{\sqrt{P}}{2\sqrt{M\sigma^2}}\right)$, where Q is the complementary cumulative distribution function of a zero-mean, unit-variance Gaussian random variable. To find the value of P for a particular SNR, we use the relation $\text{SNR} = \frac{1}{2}\left(\frac{P}{2M\sigma^2}\right)$.

Let $J^{(i)}$ be the coherent crosstalk signal at node i. Using the fact that crosstalk effects propagate towards downstream TONs, the end-to-end BER is

$$BER_{\text{end-to-end}} = \frac{1}{2}Q\left(\frac{\frac{\sqrt{P}}{2} + \sum_{i=1}^{M} J_R^{(i)}}{\sqrt{M\sigma^2}}\right) + \frac{1}{2}Q\left(\frac{\frac{\sqrt{P}}{2} - \sum_{i=1}^{M} J_R^{(i)}}{\sqrt{M\sigma^2}}\right), \tag{1}$$

where $J_R^{(i)}$ denotes the real part of $J^{(i)}$. To obtain the worst case scenario, these components are assumed to add up constructively along the light path.

Let d denote the device output at a particular TON. We want to generate an alarm when coherent crosstalk signal is large, i.e. $|d_R| > \alpha$, where α is a preset threshold. The probability of an alarm generation at a TON at a given bit time is

$$P\{\text{Device alarm}\} = Q\left(\frac{\alpha - J_R^{(i)}}{\sigma}\right) + Q\left(\frac{\alpha + J_R^{(i)}}{\sigma}\right). \tag{2}$$

Let T denote the length of an observation period in the unit of bit transmission periods, and $\tilde{B}$ denote the average BER associated with an observation period. Among all crosstalk scenarios with BER $\tilde{B}$, the one yielding the smallest total number of device alarms in an observation period has the smallest probability of a BER alarm generation, and is thus the worst case scenario for detection. We shall describe an individual scenario in terms of crosstalk signals at all TONs at all bit transmission periods. For a fixed $\tilde{B}$, deriving the worst case scenario is equivalent to finding the crosstalk signals which are the free variables in the following optimization problem.

$$\begin{array}{ll} \text{minimize} & \frac{1}{T} \times \text{expected number of device alarms in } T \text{ bits} \\ \text{subject to} & \text{average BER in } T \text{ bits} = \tilde{B}. \end{array} \quad (3)$$

For future reference, let $A(x)$ and $B(x)$ denote the device alarm probability function and the BER function whose expressions in equations 1 and 2 are presented again below.

$$B(x) = \frac{1}{2}Q\left(\frac{\frac{\sqrt{P}}{2} - x}{\sqrt{M\sigma^2}}\right) + \frac{1}{2}Q\left(\frac{\frac{\sqrt{P}}{2} + x}{\sqrt{M\sigma^2}}\right) \quad (4)$$

$$A(x) = Q\left(\frac{\alpha - x}{\sigma}\right) + Q\left(\frac{\alpha + x}{\sigma}\right). \quad (5)$$

The following list provides all notations to be used in later analysis.

SNR: The signal-to-noise ratio of the light path.

M: The number of TONs in the light path.

T: The number of bits transmitted in an observation period.

α: The threshold value for a device alarm.

β: The threshold value for a BER alarm.

$\hat{B}$: The average BER in an observation period below which a BER alarm is not generated with very high probability (see also $\tilde{B}$ and $\overline{FP}_{\text{BER}}$).

$\tilde{B}$: The average BER in an observation period above which a BER alarm is generated with very high probability (see also $\overline{FN}_{\text{BER}}$).

$\overline{FP}_{\text{BER}}$: Upper limit on FP, the probability of falsely concluding that BER $> \tilde{B}$ when in fact BER $\leq \hat{B}$.

$\overline{FN}_{\text{BER}}$: Upper limit on FN, the probability of falsely concluding that BER $\leq \tilde{B}$ when in fact BER $> \tilde{B}$.

$A(.)$: The probability function of a device alarm generation at a TON.

$B(.)$: The end-to-end BER function of a transmitted bit.

4 Analysis on the BER Guarantees

This section provides mathematical analysis leading to BER guarantees by the monitoring system described in section 3. We proceed by solving problem (3).

Let $J_i^{(k)}$ denote the real part of a coherent crosstalk signal at the i^{th} bit at TON k. Since the degradation due to crosstalk propagates along the light path, the BER of the ith bit for crosstalk signals $(J_i^{(1)}, J_i^{(2)}, \ldots, J_i^{(M)})$ is $B(\sum_{k=1}^{M} J_i^{(k)})$. The average BER in an observation period is given by $\frac{1}{T}\sum_{i=1}^{T} B\left(\sum_{k=1}^{M} J_i^{(k)}\right)$.

In each bit transmission period, we consider that a device alarm is generated if at least one TON in the light path generates a device alarm. The expected number of device alarms in an observation period is then $\sum_{i=1}^{T}\left[1 - \prod_{k=1}^{M}\left(1 - A(J_i^{(k)})\right)\right]$.

Problem (3) can be written more specifically as

$$\text{minimize} \quad \frac{1}{T}\sum_{i=1}^{T}\left[1 - \prod_{k=1}^{M}\left(1 - A(J_i^{(k)})\right)\right] \quad \text{subject to} \quad \tilde{B} = \frac{1}{T}\sum_{i=1}^{T} B\left(\sum_{k=1}^{M} J_i^{(k)}\right). \quad (6)$$

4.1 Worst Case Scenario for $M = 1$ and $T \geq 1$

Assume a single TON in the light path. Consider the case with $T = 2$. Let J_1 and J_2 denote $J_1^{(1)}$ and $J_2^{(1)}$ respectively. The optimization problem is

$$\text{minimize } \frac{1}{2}\left(A(J_1) + A(J_2)\right) \quad \text{subject to } \tilde{B} = \frac{1}{2}\left(B(J_1) + B(J_2)\right). \tag{7}$$

The Lagrange multiplier method gives us the following constraints for an extremum.

$$A'(J_i) = \lambda B'(J_i), \; i = 1, 2 \tag{8}$$

where $A'(J_i)$ is a derivative of function $A(J_i)$ with respect to J_i.

Because $A'(J_i)$ and $B'(J_i)$ are even, we shall consider only nonnegative values of J_i. Since $A'(0) = B'(0) = 0$, zero is a solution to (8). By lemma 1 which we state without proof (see [7]), under a mild assumption, there is at most one positive solution to (8).

Assumption 1 *Let σ denote the variance of both real and imaginary parts of the AWGN at any TON at any bit time, α denote the device alarm threshold, and $\eta = \sqrt{P}/2$. Assume that $\sigma < \alpha < \eta - \sqrt{3}\sigma$.*

Lemma 1 *Given $\lambda > 0$, and assumption 1, there is at most one positive solution to the equation $A'(J) = \lambda B'(J)$ in $(0, \eta - \sigma)$.*

Notice that assumption 1 generally holds when the SNR is sufficiently high since SNR $= \frac{1}{2}\left(\frac{2\eta^2}{M\sigma^2}\right)$ and therefore σ is small compared to η. *We shall assume that assumption 1 holds throughout this section,* even though empirical data suggest that our results hold regardless of the validity of assumption 1.

Given that only zero and one other positive value are solutions to (8), there are two forms of solutions for an extremum, namely $J_1 = J_2$, and J_1 (or J_2) $= 0$ with J_2 (or J_1) > 0. Let $\tilde{J}$ be such that $B(\tilde{J}) = \tilde{B}$ and J^s be such that $\frac{1}{2}B(0) + \frac{1}{2}B(J^s) = \tilde{B}$. It follows that the extrema are $(\tilde{J}, \tilde{J})$, $(0, J^s)$, and $(J^s, 0)$.

Given the locations of the extrema, we can investigate their properties by evaluating the bordered Hessian $|H|_{(J_1,J_2)}$ of the Lagrangian. By lemma 2 which we state without proof (see [7]), we have that $(\tilde{J}, \tilde{J})$ corresponds to a local maximum, and $(J^s, 0)$ and $(0, J^s)$ correspond to local minima.

Lemma 2 *Under assumption 1, $|H|_{(J^s,0)} = |H|_{(0,J^s)} < 0$ and $|H|_{(\tilde{J},\tilde{J})} > 0$, for all $\tilde{B} < \frac{1}{2}B(0) + \frac{1}{2}B(\eta - \sigma)$).*

Since the constraint set defined in (7) is a contour curve of $(B(J_1) + B(J_2))/2$ in the first quadrant of the (J_1, J_2) plane, and this curve connects the two end points $(J_s, 0)$ and $(0, J_s)$, the absence of a local minimum elsewhere guarantees that $(J^s, 0)$ and $(0, J^s)$ are indeed global minima. We conclude the result in theorem 1.

Theorem 1 *Under assumption 1, given the average degraded BER of $\tilde{B} < \frac{1}{2}B(\eta - \sigma) + \frac{1}{2}B(0)$, a scenario in which only one out of two bits ($T = 2$) is affected by coherent crosstalk at a single TON ($M = 1$) has the smallest expected number of device alarms in an observation period and is therefore the worst case scenario.*

We can extend theorem 1 to the case of $T > 2$. Let $J_i = J_i^{(1)}$. The problem is

$$\text{minimize} \quad \frac{1}{T}\sum_{i=1}^{T} A(J_i) \quad \text{subject to} \quad \tilde{B} = \frac{1}{T}\sum_{i=1}^{T} B(J_i). \tag{9}$$

The Lagrange multiplier method gives the same condition as in (8).

$$A'(J_i) = \lambda B'(J_i),\ i = 1, ..., T \tag{10}$$

We have from lemma 1 that under assumption 1 there exists at most one positive solution to $A'(J_i) = \lambda B'(J_i)$ in $[0, \eta - \sigma)$. We shall assume that $\tilde{B} < \frac{1}{T}B(\eta - \sigma) + \frac{T-1}{T}B(0)$ to guarantee that $J_i \in [0, \eta - \sigma)$ for all i in all scenarios.

Denote a scenario in which t out of T bits are affected by crosstalk signals of magnitude J_t^s (and the result average BER is $\tilde{B}$) as scenario t. All possible scenarios are 1,2,...,T. Lemma 1 tells us that an extremum corresponds to one of the T scenarios just described. Let A_t denote the expected number of device alarms in an observation period corresponding to scenario t. By lemma 3 which we state without proof (see[7]), we have that $A_T > A_{T-1} > ... > A_1$ and can conclude the result in theorem 2.

Lemma 3 *Under assumption 1, given* $\tilde{B} < \frac{1}{T}B(\eta - \sigma) + \frac{T-1}{T}B(0)$, *for scenarios with* $M = 1$ *and* $T > 1$, *we have that* $A_T > A_{T-1} > ... > A_1$.

Theorem 2 *Under assumption 1, given* $\tilde{B} < \frac{1}{T}B(\eta - \sigma) + \frac{T-1}{T}B(0)$, *a scenario in which only one out of* T *bits* ($T \geq 1$) *is affected by coherent crosstalk at a single TON* ($M = 1$) *has the smallest expected number of device alarms in an observation period and is therefore the worst case scenario.*

4.2 Worst Case Scenario for $M \geq 1$ and $T = 1$

Let $J^{(k)} = J_1^{(k)}$. Let $C(J) = 1 - A(J)$. Problem (6) becomes

$$\text{minimize} \quad -\prod_{k=1}^{M} C(J^{(k)}) \quad \text{subject to} \quad \tilde{J} = \sum_{k=1}^{M} J^{(k)}. \tag{11}$$

The Lagrange multiplier method yields the constraint for an extremum

$$-\lambda = C'(J^{(1)})\prod_{k\neq 1} C(J^{(k)}) = C'(J^{(2)})\prod_{k\neq 2} C(J^{(k)}) = ... = C'(J^{(M)})\prod_{k\neq M} C(J^{(k)}).$$

Multiplying through by $\left[\prod_{k=1}^{M} C(J^{(k)})\right]^{-1}$, we get the following relation.

$$\frac{C'(J^{(1)})}{C(J^{(1)})} = \frac{C'(J^{(2)})}{C(J^{(2)})} = ... = \frac{C'(J^{(M)})}{C(J^{(M)})}$$

Lemma 4, which we state without proof (see [7]), states that $C'(J)/C(J)$ is a strictly decreasing one-to-one function in $[0, \alpha)$.

Lemma 4 *The function* $C'(J)/C(J)$ *strictly decreases in* $[0, \alpha)$.

It follows that for any value of $\tilde{J} < M\alpha$, an extremum must have equal components, namely $(\tilde{J}/M, ..., \tilde{J}/M)$, and is thus unique. To show that $(\tilde{J}/M, ..., \tilde{J}/M)$

is indeed the global minimum, it is sufficient to show that the cost function at another point in the constraint set is larger than at the extremum. Consider the point $(3\tilde{J}/M, -\tilde{J}/M, \tilde{J}/M, ..., \tilde{J}/M)$, we have that

$$-C(3\tilde{J}/M)C(-\tilde{J}/M)C(\tilde{J}/M)^{M-2} = -C(3\tilde{J}/M)C(\tilde{J}/M)^{M-1} > -(C(\tilde{J}/M))^M,$$

where the equality follows from the fact that $A(J)$ is symmetric; and so is $C(J) = 1 - A(J)$. The inequality follows from the fact that $C(3\tilde{J}/2) < C(\tilde{J}/2)$ since $C(J)$ is a strictly decreasing function (its first derivative $-A'(J)$ is negative in $[0, \infty)$). We have thus established theorem 3.

Theorem 3 *Given $\tilde{B} < B(M\alpha)$, a scenario in which the crosstalk signals at M TONs ($M \geq 1$) are equal has the smallest expected number of device alarms in an observation period ($T = 1$) and is therefore the worst case scenario.*

4.3 Worst Case Scenario for $M \geq 1$ and $T \geq 1$

As before, let $C(J) = 1 - A(J)$. Problem (6) can be expressed as

$$\text{minimize} \quad -\frac{1}{T}\sum_{i=1}^{T}\prod_{k=1}^{M} C(J_i^{(k)}) \quad \text{subject to} \quad \tilde{B} = \frac{1}{T}\sum_{i=1}^{T} B\left(\sum_{k=1}^{M} J_i^{(k)}\right). \tag{12}$$

In any optimal solution, we must have, for each i, the minimum value of $-\prod_{k=1}^{M} C(J_i^{(k)})$ subject the constraint on $\sum_{k=1}^{M} J_i^{(k)}$. Otherwise, we can reassign the values of $J_i^{(k)}$ such that their sum remains the same but $-\prod_{k=1}^{M} C(J_i^{(k)})$ is smaller.

Let J_i denote $\sum_{k=1}^{M} J_i^{(k)}$. Let $C_M(J_i)$ denote the maximal value of $\prod_{k=1}^{M} C(J_i^{(k)})$ for the ith bit interval subject to the constraint on $\sum_{k=1}^{M} J_i^{(k)}$. In terms of $C_M(J)$, problem (12) can be expressed as

$$\text{minimize} \quad -\frac{1}{T}\sum_{i=1}^{T} C_M(J_i) \quad \text{subject to} \quad \tilde{B} = \frac{1}{T}\sum_{i=1}^{T} B(J_i). \tag{13}$$

Comparing (13) with (9), we can consider problem (13) as if we were to find the worst case scenario for $M = 1$ and $T \geq 1$, but with the function $-C_M(J)$ instead of $A(J)$ as in section 4.1.

Consider the function $C_M(J)$. For $J < M\alpha$, theorem 3 tells us that $C_M(J) = (C(J/M))^M$ since in the worst case scenario, coherent crosstalk signals at M TONs are equal. Using the same approach as in section 4.1, the Lagrange multiplier method yields the constraint for an extremum

$$C'_M(J_i) = \lambda B'(J_i), \; i = 1, ..., T. \tag{14}$$

Empirical data suggests the similar result as in section 4.1, namely the curve $-C'_M(J)$ intersects $B(J)$ at zero and one other positive value (for $J \geq 0$), and the extrema correspond to scenarios in which some bits are not degraded while the others are degraded equally (denoted as scenarios 1 to T in section 4.1). However, we have no theoretical proof of this claim. The difficulty lies in the complexity of an expression for $-C'_M(J)$. Numerical verification in several cases leads us to the following conjecture.

Given $\tilde{B} < \frac{1}{T}B(M\alpha) + \frac{T-1}{T}B(0)$, the worst case scenario ($M \geq 1, T \geq 1$) is the one in which: (*i*) only one out of T bits in the observation period is affected by coherent crosstalk signals, (*ii*) the same amount of crosstalk signals are present at all M TONs in the light path.

5 BER Monitoring System with Hard Limiters

For a low guaranteed BER, under the worst case scenario found in section 4, a single severely degraded bit would only generate a single extra device alarm (compared to the case with no degradation) which leads to high values for FP and FN probabilities. Therefore, we propose the use of hard limiters at TON inputs to improve the performance of the BER monitoring system.

A hard limiter is a device that limits the power of signals passing through it. For crosstalk signals to degrade few bits significantly, the instantaneous crosstalk signal power in those bit intervals must be very high. But very high instantaneous crosstalk signal power results from very high instantaneous input power, which are not permitted in the presence of hard limiters at TON inputs. Moreover, since the hard limit is generally higher than the ON level signal power of OOK, the use of hard limiters does not interfere with normal operation.

In the presence of hard limiters, multiple bits must be degraded for the significant amount of BER degradation. Given that at least t ($t > 1$) bits must be degraded, the worst case scenario is the one in which t out of T bits are equally affected and crosstalk signals at M nodes are equal. When the hard limit is sufficiently low (the value of t is sufficiently high), our BER monitoring system will be able to detect the worst case scenario for a particular guaranteed BER. We shall demonstrate with examples.

Example 1 Assume the following parameters: SNR = 20 dB, M=1, $\hat{B}=10^{-12}$, $\tilde{B}=10^{-8}$, $\overline{FP}_{\text{BER}} = 10^{-8}$, $\overline{FN}_{\text{BER}} = 10^{-8}$, and a TON crosstalk level of -30 dB.

With a hard limit of $60P$, $\alpha = 0.19\sqrt{P}$, and $\beta = 329$, the system is able to detect a BER above 10^{-8} in $1\mu s$ ($T = 1,000$).

Example 2 Assume the same parameters as in example 1 but with $M = 5$ instead of $M = 1$.

With a hard limit of $2P$, $\alpha = 0.05\sqrt{P}$, and $\beta = 827$, the system is able to detect a BER above 10^{-8} in $1\mu s$ ($T = 1,000$).

Note that the cut-off level decreases as M increases. This observation suggests that there is a limit on the number of TONs in a light path above which the cut-off level of hard limiters gets too close to the power level under normal operation of TONs.

6 Conclusion and Future Directions

We derived guaranteed BERs provided by the novel BER monitoring system in [1, 2]. In doing so, we found that the worst case coherent crosstalk scenario for the BER monitoring system to detect is the one in which only one out of T bits in the observation period is affected by coherent crosstalk signals of the same amount at all TONs in the light path. Based on this result, we suggested the use of hard limiters at TON inputs to improve the system performance.

The BER monitoring system considered in this work offers several advantages. The system can detect BER degradation much faster than a BERT (6 orders of magnitude faster in our examples). The service monitoring devices can be installed at existing TONs so no modification of existing infrastructure is required. Although our analysis is based on OOK signalling, the same method can be applied in the cases of other transmission schemes as well. Finally, since the system does not differentiate between the sources of signal degradation at TONs, it can be used to detect malicious users who intentionally cause service disruption.

Several related issues require further consideration. First, a similar kind of analysis to this work can be done on BER degradation due to crosstalk effects as well as gain competition at EDFAs. Another issue has to do with actual implementation of the BER monitoring system. While the installation of service monitoring devices at TONs to monitor different WDM light paths seems straight forward in a WDM network with static wavelength routing, the implementation becomes complex if we allow dynamic wavelength routing. Finally, there are also issues regarding how to provide appropriate reponses after the detection of BER degradation.

References

[1] M. Médard, S.R. Chinn, P. Saengudomlert, "Attack Detection in All-Optical Networks," *OFC'98*, ThD4.

[2] M. Médard, S.R. Chinn, P. Saengudomlert, "Attack and Failure Detection in Transparent Optical Nodes," Lincoln Laboratory internal report, 1998.

[3] L.A. Buckman, L.P. Chen, K.Y. Lau, "Crosstalk Penalty in All-Optical Distributed Switching Networks," *IEEE Photonics Technology Letters*, vol. 9, no. 2, February 1997.

[4] C.S. Li, F. Tong, "Crosstalk and Interference Penalty in All-Optical Networks Using Static Wavelength Routers," *Journal of Lightwave Technology*, vol. 14, no. 6, June 1996.

[5] C. Saxtoft, P. Chidgey, "Error Rate Degradation Due to Switch Crosstalk in Large Modular Switched Optical Networks," *IEEE Photonics Technology Letters*, vol. 5, no. 7, July 1993.

[6] C.X. Yu, W.K. Wang, S.D. Brorson, "System Degradation Due to Coherent Crosstalk in WDM Network Nodes," *OFC'98*, WM 30.

[7] P. Saengudomlert, "Analysis and Detection of Jamming Attacks in an All-Optical Network," MIT Master Thesis, 1998.

[8] G.R. Hill et al, "A Transport Network Layer Based on Optical Network Elements," *Journal of Lightwave Technology*, vol. 11, no. 5/6, May/June 1993.

[9] A. Kloch, B. Mikkelsen, K.E. Stubkjaer, "Pilot tones in WDM networks with wavelength converters," *OFC'97*, TuE6.

A Fast Special Factorization Algorithm in the Sudan Decoding Procedure

Gui-liang Feng[*]

The Center for Advanced Computer Studies
University of Louisiana at Lafayette
Lafayette, LA 70504
glf@cacs.usl.edu

Abstract

In the Sudan decoding procedure, there are two stages: weighted curve fitting and factorization of a bivaricate polynomial. In this paper, we present a fast algorithm with complexity $O(s^3 n \log^2 n)$ for the special factorization of a bivariate polynomial where s is the degree of y in $Q(x, y)$. Combining this fast algorithm with a fast generalized Welch-Berlekamp algorithm, the complexity of the Sudan decoding procedure for Reed-Solomon codes can be reduced to $O(n \log^3 n)$.

1 Introduction

Let $GF(q)\underline{\underline{\triangle}}\{0, 1, a, a^2, \cdots, a^{q-2}\}$ be a finite field, where a is a primitive element of $GF(q)$. Let $x_1, x_2, \cdots, x_n$ be n distinct elements of $GF(q)$. Usually, we let $n = q$. We difine a Reed-Solomon code as the following:

$$C = \{c = (c_1, c_2, \cdots, c_n \underline{\underline{\triangle}}$$

$$\underline{\underline{\triangle}}(C(x_1), C(x_2), \cdots, C(x_n))\ |C(x) \in GF(q)[x] and \quad \deg C(x) < k < n\}.$$

When $n = q$, C is called an extended Reed-Solomon code. This is a $[n, k, d]$ code over $GF(q)$, where n is the codelength, k is the code information bit, and d is the minimum distance. We have $d = n - k + 1$. Let $t^* = \lfloor \frac{d-1}{2} \rfloor$,and t^* is the capability of correcting errors.

Let $y\underline{\underline{\triangle}}(y_1, y_2, \cdots, y_n)$be the received vector, and let $[n]$ denote the set of integers $\{1, \cdots, n\}$. The Sudan decoding procedure consists of finding all polynomials $C(x)$ within the Hamming distance τ of the received vector y, i.e., $c_i = y_i$ for $(n - \tau)$ positions, where τ is an integer greater than or equal to t^*. The Sudan decoding procedure can correct up to τ errors and τ can be greater than t^* in the case of low rate code.

By saying that the algorithm can correct up to τ errors, or the error-correcting capability is τ, we mean that the number of errors is less than or equal to τ. The algorithm can find all codewords $C(x) \in GF(q)[(x)]$ within distance τ of the received vector. Then the side information, like the soft channel information, may be used to help determine which codeword is most likely to be the transmitted one. In order to describe the Sudan decoding procedure the following definition is needed.

---

[*] This work was supported in part by the NSF under Grant NCR-9804973

Definition 1.1 *Let the (w_x, w_y)-weighted degree of a monomial x^iy^j be $iw_x + jw_y$, denoted by* $\mathrm{Deg}[\mathrm{x^iy^j}]$. *Let the (w_x, w_y)-weighted degree of polynomial* $\mathrm{Q(x,y)} \triangleq \sum_{\mathrm{i}}\sum_{\mathrm{j}} \mathrm{q_{ij}x^iy^j}$ *be* $\max_{q_{ij}\neq 0} Deg[x^iy^j]$, *denoted by $DegQ(x,y)$.*

The Sudan decoding procedure is based on the following theorem, where $w_x = 1$ and $w_y = k - 1$ [1-5].

Theorem 1.1 *If a polynomial* $\mathrm{Q(x,y)} \in \mathrm{GF(q)[x,y]}$ *satisfies.*

$$DegQ(x,y) < n - \tau \tag{1.1}$$

and

$$Q(x_i, y_i) = 0 \qquad \forall i \in [n] \quad and \quad Q(x,y) \quad is \quad not \quad identically \quad zero, \tag{1.2}$$

then $\mathrm{(y - C(x))}$ *is a factor of* $\mathrm{Q(x,y)}$ *for all codewords* $\mathrm{C(x)}$ *within the Hamming distance τ of the received vector.*

The Sudan decoding procedure involves two major stages:(1) to find $Q(x,y)$ such that (1.1) and (1.2) are satisfied (it is known as the weighted curve fitting problem); and (2) to find all factors $(y - C(x))$ of $Q(x,y)$ (it is known as the special factorization problem). In this paper, we consider the second problem only.

Let $Q(x,y)$ be a bivariate polynomial. The degree of y is s, i.e.,

$$Q(x,y) = \sum_{i=0}^{s} p_i(x)y^i, \quad where \quad p_s(x) \neq 0.$$

The special factorization problem is to find all $y(x)$ such that

$$Q(x, y(x)) = \sum_{i=0}^{s} p_i(x)y^i(x) \equiv 0 \tag{1.3}$$

In the next section, we present an algorithm for the special fatctorization problem with complexity $O(s^3n^2)$, where s is the degree of y. Then in Section III, applying the divide-and-conquer technique, we derive a divide-and-conquer algorithm to solve the special factorization problem with complexity $O(s^3 n \log^2 n)$. Finally, we present our conclusions in Section IV.

2 An algorithm for the Special Factorization Problem

In this section, we consider the special factorization problem: given $Q(x,y) = \sum_{i=0}^{s} P_i(x)y^i$, where $P_i(0) \neq 0$, find all possible $Y(x)$ such that

$$Q(x, Y(x)) = \sum_{i=0}^{s} P_i(x)Y^i(x) \equiv 0,$$

where n is the $\max\{\deg P_i(x) \mid 0 \leq i \leq s\}$. First, we introduce some useful lemmas:

Lemma 2.1 *Let* $A(x) = \sum_{i=0}^{m} a_i x^i$, $B(x) = \sum_{i=0}^{m} b_i x^i$, *and* $Y(x) = \sum_{i=0}^{m} y_i x^i$. *Let us consider the equation:*

$$A(x)Y(x) + B(x) = 0, \tag{2.1}$$

where $a_0 \neq 0$, *we have*

$$a_0 y^i + \sum_{\lambda=1}^{i} a_\lambda y_{i-\lambda} + b_i = 0 \quad for \quad i = 0, 1, \cdots, m. \tag{2.2}$$

Lemma 2.2 *Let* $A(x, y)$ *be a polynomial in* x, y, $Y(x) = \sum_{i=0}^{m} y_i x^i$, *and* $A^{(j)} = \sum_{i=0}^{m} a_i^{(j)}$ *for* $j = 0, 1, \cdots, p$. *Let*

$$F(x) \triangleq A(x, Y(x)) x Y^{p+1}(x) + \sum_{j=0}^{p} A^{(j)}(x) Y^j(x),$$

where $a_0^{(p)} \neq 0$, *then we have*

$$\sum_{j=0}^{p} a_0^{(j)} y_0^j = 0. \tag{2.3}$$

Lemma 2.3 *Let us consider the equation over a field* F*:*

$$E(y) = \sum_{j=0}^{p} a^{(j)} y^j = 0 \qquad a^{(0)} \neq 0. \tag{2.4}$$

Suppose that y^* *is a* p*-multiplicity root of (2.4). If* p *is not divisible by the character* q *of* F *then* $y^* = a^{(p-1)}/pa^{(p)}$, *otherwise* $y^* = \left(a^{(q^u(h-1))}/a^{(p)}h\right)^{-q^u}$, *where* $p = hq^u$.

Proof: Since y^* is a p-multiplicity root of (2.4), we have $E(y) = a^{(p)}(y - y^*)^p$. If p is not divisible by the character of F, we have

$$E(y) = a^{(p)} y^p + a^{(p)} p y^{p-1} y^* + \cdots.$$

Comparing the coefficient of y^{p-1} in the above equation to (2.4), we have $a^{(p-1)} = a^{(p)} p y^*$, i.e., $y^* = a^{(p-1)}/pa^{(p)}$.

If $p = hq^u$ for $\mu \geq 1$, then we have

$$E(y) = a^{(p)}(y^{q^\mu} - y^{*q^\mu})^h = a^{(p)}(y^{q^\mu})^h + a^{(p)} h y^{*q^\mu} (y^{q^\mu})^{h-1} + \cdots$$

Comparing the coefficient of $(y^{q^\mu})^{h-1}$ in the above equation to (2.4), we have

$$a^{(q^\mu(h-1))} = a^{(p)} h y^{*q^\mu}, \quad i.e., \quad y^* = \left(a^{(q^\mu(h-1))}/a^{(p)}h\right)^{-q^\mu}.$$

Corollary 2.1 *Let* $y^* = a^{(p-1)}/pa^{(p)}$ *or* $y^* = \left(a^{(q^\mu(h-1))}/a^{(p)}h\right)^{-q^\mu}$ *for* q *not divisible by* p *or* $p = hq^\mu$ *for* $\mu \geq 1$, *respectively, if* y^* *is not a root, then* $E(y) = 0$ *has no p-multiplicity root.*

Lemma 2.4 *Let* $f(y) = \sum_{i=0}^{n} p_i y^i$, y_0 *is a r-multiplicity root of* $f(y)$ *if and only if*

$$\sum_{i=0}^{n-\lambda} \binom{\lambda + i}{\lambda} p_{i+\lambda} y_0^i = 0 \quad for \quad \lambda = 0, 1, \cdots, r-1, \qquad \sum_{i=0}^{n-r} \binom{r+i}{r} p_{i+r} y_0^i \neq 0 \tag{2.5}$$

Proof: $f(y)$ can be rewritten as $f(y) = \sum_{i=0}^{n} p_i(y - y_0 + y_0)^i$. It can be extended to

$$f(y) = \sum_{i=0}^{n} p_i \sum_{\mu=0}^{i} \begin{pmatrix} i \\ \mu \end{pmatrix} (y - y_0)^{\mu} {y_0}^{i-\mu} = \sum_{i=0}^{n} (y - y_0)^{\mu} \sum_{j=0}^{n-\lambda} \begin{pmatrix} j+\lambda \\ \lambda \end{pmatrix} P_{j+\lambda} {y_0}^{j}.$$

If y_0 is a r-multiplicity root of $f(y)$ if and only if the coefficients of $(y - y_0)^{\lambda}$ are zero for $\lambda < r$ and the coefficient of $(y - y_0)^r$ is non-zero.

Let $Q(x, Y(x)) = x(\sum_{j=p+1}^{s} A_j(x)Y^j(x)) + \sum_{i=0}^{p} A_i(x)Y^j(x)$ and $a_{p,0} \neq 0$, where $A_{\mu}(x) = \sum_{\lambda=0} a_{\mu,\lambda}x^{\lambda}$, then we call $Q(x, Y(x)) = 0$ a quasi-p degree equation.

Lemma 2.5 *A Quasi-p degree equation in $Y(x)$ can be solved by the algorithm for a p degree equation in $Y(x)$ in the order of $y_0, y_1, \cdots, y_m$.*

Proof: By the algorithm for a p degree equation in $Y(x)$ in the order of $y_0, y_1, \cdots, y_m$, the process of finding y_0 for the p degree equation and for the corresponding quasi-p degree equation is the same. Assume we have found $y_0, \cdots, y_i$ for these two equations, next we find y_{i+1}. For the corresponding quasi-p degree equation, we should calculate the coefficient of x^{i+1} in $x(\sum_{j=p+1}^{s} A_j(x)Y^j(x))$ and add it to the coefficient of x^{i+1} in $A_0(x)$; then, follow the same approach for the p degree equation.

2.1 The Case of $s = 1$:

First, we consider the simplest case of $s = 1$, i.e., $Q(x, y) = A(x)y + B(x)$, where $A(x) = \sum_{i=0}^{m} a_i x^i$ and $B(x) = \sum_{i=0}^{2m} b_i x^i$. The special factorization problem is to find $Y(x) = \sum_{i=0}^{m} y_i x^i$ from the following equation:

$$A(x)Y(x) + B(x) = 0. \tag{2.7}$$

Without loss of generality, we can assume $a_0 \neq 0$. Using Lemma 2.1, we can find y_0, $y1$,$\cdots$,y_m. Then, we should verify by checking that the equation

$$\sum_{j=0}^{m} a_j y_{\lambda+m-j} + b_{\lambda+m} = 0 \qquad for \qquad \lambda = 1, 2, \cdots, m,$$

where $y_j = 0$ for $j > m$. The complexity is $O(m^2)$. Using the divide-and-conquer technique, the complexity can be reduced to $O(m\log^2 m)$.

The algorithm based on Lemma 2.1 can be described as a recurrent procedure which can be generalized to the case of $s > 1$. We need to introduce some new notation.

Let $F(x) = \sum_{i=0}^{m} f_i x^i$ be a polynomial. We define

$$F(x) = F_0(x) = f_0 + F_1(x)x = \cdots = \sum_{i=0}^{\mu-1} f_i x^i + F_{\mu}(x)x^{\mu}.$$

Let $A^{(0)}(x) = A(x)$ and $B^{(0)}(x) = B(x)$, where the superscript denotes the iterative number. Thus, (2.7) can be written as

$$A^{(0)}(x)Y_0(x) + B^{(0)}(x) = 0. \tag{2.7$'$}$$

It is equal to

$$A^{(0)}(x)xY_1(x)+(A_1^{(0)}(x)y_0+B_1^{(0)}(x))x+a_0^{(0)}y_0+b_0^{(0)}=0.$$

By Lemma 2.2, we have $a_0^{(0)}y_0+b_0^{(0)}=0$ and find $y_0=-b_0^{(0)}/{a_0}^{(0)}$. Let $A^{(1)}(x)=A^{(0)}(x)$ and $B^{(1)}=A_1^{(0)}(x)y_0+B_1^{(0)}(x)$, then (2.7') reduce to

$$A^{(1)}(x)Y_1(x)+B^{(1)}(x)=0. \tag{2.8}$$

Similarly, from (2.8) we can find $y_1=-b_0^{(1)}/{a_0}^{(1)}$ and (2.8) reduces to

$$A^{(2)}(x)Y_2(x)+B^{(2)}(x)=0. \tag{2.9}$$

when $A^{(2)}(x)=A^{(1)}(x)$ and $B^{(2)}=A_1^{(1)}(x)y_1+B_1^{(1)}(x).=0$. The recurrent procedure is the following. If $y_{\mu-1}$, $A^{(\mu)}(x)$ and $B^{(\mu)}(x)$, for $\mu<m$, are found, then we have

$$A^{(\mu)}(x)Y_\mu(x)+B^{(\mu)}(x)=0. \tag{2.10}$$

From (2.10), we can find $y_\mu=-b_0^{(\mu)}/a_0^{(\mu)}$, $A^{(\mu+1)}(x)=A^{(\mu)}(x)$ and $B^{(\mu+1)}(x)=A_1^{(\mu)}(x)y_1$ $+B_1^{(\mu)}(x)$. Once $y_0,\cdots,y_m$ are found, then we should verify by checking the m equations. At the μth recurrent step, we need μ multiplications and one division. Thus, the complexity is $O(m^2)$. Using the divide- and-conquer technique, the complexity can be reduced to $O(m\log^2 m)$.

2.2 The case of $s=2$:

Let $A(x)=\sum\limits_{i=0}^{m}a_ix^i$, $B(x)=\sum\limits_{i=0}^{2m}b_ix^i$, $C(x)=\sum\limits_{i=0}^{3m}c_ix^i$, where a_0 and b_0 are not all zero, and let $Q(x,y)=A(x)y^2+B(x)y+C(x)$. We are going to find $Y(x)=\sum\limits_{i=0}^{m}y_ix^i$ such that

$$Q(x,Y(x))=A(x)Y^2(x)+B(x)Y(x)+C(x)=0. \tag{2.11}$$

Let $A^{(0)}=A(x)$,$B^{(0)}=B(x)$,and $C^{(0)}(x)=C(x)$,(2.11) is equal to

$$A^{(0)}(x)(xY_1(x)+y_0)^2+B^{(0)}(x)(xY_1(x)+y_0)+C^{(0)}(x)=0.$$

$$\begin{aligned}x[A^{(0)}(x)xY_1^2(x)+(2A^{(0)}(x)y_0+B^{(0)}(x))Y_1(x)+(A_1^{(0)}(x)y_0^2+\\ B_1^{(0)}(x)y_0+C_1^{(0)}(x))]+a_0^{(0)}y_0^2+b_0^{(0)}y_0+c_0^{(0)}=0.\end{aligned} \tag{2.12}$$

By Lemma 2.2, we have

$$a_0^{(0)}y_0^2+b_0^{(0)}y_0+c_0=0. \tag{2.13}$$

Without loss of generality, we assume $a_0\neq 0$. For (2.13), there are three cases: no root, two distinct roots, and one double root. By Lemma 2.3, we test if y^* is a root of (2.13). If y^* is a root of (2.13), then y^* is the double root; otherwise, we should solve (2.13). If there is no solution for (2.13), then (2.11) has no solution $Y(x)$. Otherwise, we can find two distinct roots $y_0^{'}$ and $y_0^{''}$. Consider the following cases.

Case 1: $y_0^{'}$ and $y_0^{''}$ are simple roots. Then for each one, we want to find the corresponding $Y(x)$. For example, we consider $Y(x)$ corresponding to $y_0^{'}$, i.e., $y_0=y_0^{'}$.

Since $y_0^{'}$ is a root of (2.13), (2.12) reduces to

$$A^{(0)}(x)xY_1^2(x)+(2A^{(0)}(x)y_0^{'}+B^{(0)}(x))Y_1(x)+(A_1^{(0)}(x)y_0^{'2}+B_1^{(0)}(x)y_0^{'}+C_1^{(0)}(x))=0. \tag{2.14}$$

Let $A^{(1)}(x) = A^{(0)}(x)$,$B^{(1)}(x) = 2A^{(0)}(x)y_0 + B^{(0)}(x)$ and $C^{(1)}(x) = A_1^{(0)}(x)y_0'^2 + B_1^{(0)}(x)y_0' + C_1^{(0)}(x)$, then (2.14) can be written as

$$A^{(1)}(x)xY_1^2(x) + B^{(1)}(x)Y_1(x) + C^{(1)}(x) = 0. \tag{2.14'}$$

Since y_0' is a simple root of (2.13), $2a_0^{(0)}y_0' + b_0^0 \neq 0$. Thus, $b_0^{(1)} = 2a_0^{(0)}y_0' + b_0^0 \neq 0$. By Lemma 2.5, (2.14') is equivalent to a linear equation about $Y_1(x)$. We can use the algorithm in the case of $s = 1$ to find $y_1', \cdots, y_m'$, i.e., $Y'(x)$. We use the same algorithm to find $Y''(x)$.

Case 2: y_0^* is the double root of (2.13).
In this case (2.12) reduces to

$$A^{(0)}(x)xY_1^2(x) + (2A_1^{(0)}(x)y_0^* + B_1^{(0)}(x))Y_1(x) + (2a_0^{(0)}y_0^* + b_0^{(0)})Y_1(x) + A_2^{(0)}(x)(y_0^{*2}) + \\ B_2^{(0)}(x)y_0{}^* + C_2^{(0)}(x))x + a_1^{(0)}(y_0^*)^2 + b_1^{(0)}y_0^* + c_1^{(0)} = 0. \tag{2.15}$$

Since y_0^* is the double root of (2.13), we have $2a_0^{(0)}y_0' + b_0^{(0)} = 0$ and (2.12) reduces to

$$A^{(0)}(x)xY_1^2(x) + (2A^{(0)}(x)y_0^* + B_1^{(0)}(x))Y_1(x) + (A_2^{(0)}(x)(y_0^{*2}) + B_2^{(0)}(x)y_0{}^* + C_2^{(0)}(x))x + \\ a_1^{(0)}(y_0^*)^2 + b_1^{(0)}y_0^* + c_1^{(0)} = 0. \tag{2.15'}$$

If (2.15') has a solution, it is necessary that

$$a_1^{(0)}(y_0^*)^2 + b_1^{(0)}y_0^* + c_1^{(0)} = 0. \tag{2.16}$$

If (2.16) is not satisfied, then (2.12) has no solution, otherwise (2.15') reduces to

$$A^{(1)}(x)xY_1^2(x) + B^{(1)}(x)Y_1(x) + C^{(1)}(x) = 0. \tag{2.17}$$

where $A^{(1)}(x) = A^{(0)}(x)$, $B^{(1)}(x) = 2A_1^{(0)}(x)y_0^* + B_1^{(0)}(x)$ and $C^{(1)}(x) = A_2^{(0)}(x)y_0^{*2} + B_2^{(0)}(x)y_0^* + C_2^{(0)}(x)$.

It can be easily seen that (2.17) is similar to (2.11). We can use the same approach that is used for (2.11) to solve (2.17). Repeat this recurrent step; if there is a solution, we can find all solutions. We should mention that, in this recurrent step, we solve (2.13) only one time. Assume that the complexity is less than $O(m)$. Thus, the total complexity is $O(m^2)$. If using the divide-and-conquer technique, the complexity can be reduced to $O(m\log^2 m)$.

2.3 The General Case of $s \geq 3$:

Now we are going to describe the algorithm for the general case. Let $P^\mu(x) \overset{\triangle}{=} \sum\limits_{i=0}^{(s-\mu+1)m} p_i^\mu x^i$ for $\mu = 0, 1, \cdots, s$, and we assume $p_0^{(s)} \neq 0$. Let $Q(x,y) = \sum\limits_{\mu=0}^{s} P^{(\mu)}(x)y^\mu$; then, we consider the following equation:

$$Q(x, Y(x)) = \sum_{\mu=0}^{s} P^{(\mu)}(x)Y^\mu(x) = 0. \tag{2.18}$$

As in the cases of $s = 1$ and $s = 2$, we would like to find all possible solutions $Y(x)$. Using a recurrence algorithm, we will add an superscript to note the recurrent steps. Let

$P^{(\mu,0)}(x) = \sum_{i=0}^{(s-\mu)m} p_i^{(\mu,0)} x^i \triangleq P_i^{(\mu,0)}(x)$. The superscript 0 denotes the initial step. Let $P^{(*,0)}(x, Y_0(x)) = 0$. Thus, (2.18) reduces to

$$P^{(*,0)}(x, Y_0(x)) Y_0^{s+1}(x) x + \sum_{\mu=0}^{s} p_i^{(\mu,0)}(x) Y_0^{\mu}(x) = 0. \tag{2.18'}$$

It is a quasi-s degree equation and is equivalent to

$$\sum_{\mu=0}^{s} p_i^{(\mu,0)}(x)(xY_1(x) + y_0)^{\mu} = \sum_{\mu=0}^{s} p_i^{(\mu,0)}(x) \sum_{l=0}^{\mu} \binom{\mu}{l} x^l Y_1^l(x) y_0^{\mu-l} = 0.$$

It can be written as

$$x(\sum_{l=1}^{s} x^{l-1}(\sum_{\mu=l}^{s} \binom{\mu}{l} y_0^{\mu-l} p^{(\mu,0)}(x)) Y_1^l(x) + \sum_{\mu=0}^{s} p_1^{(\mu,0)}(x) y_0^{\mu}) + \sum_{\mu=0}^{s} p_0^{(\mu,0)} y_0^{\mu} = 0. \tag{2.19}$$

To find the $Y(x)$, we should first find y_0. From (2.19), we have

$$\sum_{\mu=0}^{s} p_0^{(\mu,0)} y_0^{\mu} = 0. \tag{2.20}$$

Using Corollary 2.1, we can easily check whether y_0^* is the s-multiplicity root. We have two cases: yes or no. If y_0^* is the s-multiplicity root, we need to solve (2.20); otherwise, we need to solve (2.20), i.e., a s-degree equation. But any root is a δ-multiplicity root for $\delta < s$.

Let y_0^* be a r-multiplicity root of (2.20), where $0 < r \leq s$. Let us consider the left side of (2.18'). It is equal to

$$\begin{gathered}\sum_{k=r+1}^{s} (p_{r+1}^{(k,0)}(x) x^{r+1} + \sum_{\lambda=0}^{r} p_{\lambda}^{(k,0)} x^{\lambda})(\sum_{l=0}^{k} \binom{k}{l} x^l Y_1^l(x)(y_0^*)^{k-l}) + \\ \sum_{\mu=0}^{r} (p_{r-\mu}^{(\mu,0)}(x) x^{r-\mu} + \sum_{\lambda=0}^{r-\mu-1} p_{\lambda}^{(\mu,0)} x^{\lambda})(\sum_{l=0}^{\mu} \binom{\mu}{l} x^l Y_1^l(x)(y_0^*)^{\mu-l}).\end{gathered} \tag{2.21}$$

Regarding the above $p_{r+1}^{(k,0)}(x)x^{r+1}$ and $p_{r+1}^{(k,0)}(x)x^{r+1}$ as terms $P_{r+1}^{(k,0)}$ and $P_{r-\mu}^{(\mu,0)}$ instead of polynomials in x, we expand (2.21). Consequently, (2.21) is equal to

$$\sum_{i=0}^{r+1} \mathbf{A}_i x^i, \tag{2.22}$$

where

$$\mathbf{A}_0 = \sum_{\mu=0}^{s} (p_0^{(\mu,0)} (y_0^*)^{\mu}); \tag{2.22'}$$

for $1 \leq j \leq r$,

$$\begin{gathered}\mathbf{A}_j = \sum_{k=r+1}^{s} \sum_{l=0}^{j} \binom{k}{l} Y_1^l(x)(y_0^*)^{k-l} p_{j-l}^{(k,0)} + \\ + \sum_{\lambda=r-j}^{r} \binom{\lambda}{j-r-\lambda} Y_1^{j-r+\lambda}(x)(y_0^*)^{r-j} \mathbf{P}_{r-\lambda}^{(\lambda,0)} + \sum_{\mu=0}^{r-j-1} \sum_{l=0}^{j} \binom{\mu}{l} Y_1^l(x)(y_0^*)^{\mu-l} p_{j-l}^{(\mu,0)}\end{gathered} \tag{2.22''}$$

$$\mathbf{A}_{r+1} = \sum_{k=r+1}^{s} [\mathbf{P}_{r+1}^{(k,0)}(xY_1(x) + y_0^*)^k + \sum_{\substack{l+\lambda \geq r+1 \\ 0 \leq \lambda \leq r \\ 0 \leq l \leq k}} \binom{k}{l} Y_1^l(x) x^{l+\lambda-r-1} (y_0^*)^{k-l} P_{\lambda}^{(k,0)}]. \tag{2.22'''}$$

Since $y_0^* = 0$, $\mathbf{A}_0$ should be zero. Now let us consider the coefficient of $Y_1^r(x)$ in $\mathbf{A}_r$. From (2.22') with $j = r$, in the first part of $\mathbf{A}_r$ there is

$$\sum_{k=r+1}^{s} \binom{k}{r} (y_0^*)^{k-r} p_{r-r}^{(k,0)} Y_1^r(x);$$

in the second part there is $\binom{r}{r} \mathbf{P}_0^{(r,0)} Y_1^r(x)$; and in the third part, there is no term $Y_1^r(x)$. Therefore, in A_r there is

$$[\sum_{k=r+1}^{s} \binom{k}{r} (y_0^*)^{k-r} p_0^{(k,0)} + \binom{r}{r} \mathbf{P}_0^{(r,0)}] Y_1^r(x),$$

i.e.,

$$[\sum_{k=r}^{s} \binom{k}{r} (y_0^*)^{k-r} p_0^{(k,0)} + P_1^{(r,0)}(x)x] Y_1^r(x),$$

From Lemma 2.4 and the fact that y_0^* is a r-multiplicity root of (2.20), we have $\sum_{k=r}^{s} \binom{k}{r} (y_0^*)^{k-r} p_0^{(k,0)} \neq 0$ and we denote it by c. Thus, the coefficient of $Y_1^r(x)$ in $\mathbf{A}_r$ is $p_1^{(r,0)}(x)x + c$, $c \neq 0$, and it is not a non-zero polynomial. Let h be the smallest integer such that in $\mathbf{A}_h$ there is a non-zero polynomial $c(x)Y_1^t(x)$ where $1 \leq t \leq h$ (in fact, if $\mathbf{A}_j$ $(j \leq r)$ contains $Y_1^i(x)$, then $i \leq j$); then, $h \leq r$ and there is such an h. Thus, in $\mathbf{A}_j$ for $j \leq h$, there is no $Y_1^i(x)$, and any polynomial in $\mathbf{A}_j$ should be zero; otherwise, the (2.18') has no solution with $y_0 = y_0^*$. Therefore, we can divide $\sum_{k=h}^{s} \mathbf{A}_k x^k$ by x^h and have

$$(\sum_{k=h+1}^{s} \mathbf{A}_k x^{k-h})x + \mathbf{A}_h = 0, \tag{2.23}$$

where $\mathbf{A}_h = \sum_{i=0}^{t} C_i(x) Y_1^i(x)$. From Lemma 2.5, (2.23) is a quasi-t degree equation, where $t \leq t$. We denote $C_i(x)$ and $\sum_{k=h+1}^{s} \mathbf{A}_k x^{k-h}$ by $P_{(i)}(x, Y_1(x)) Y_1^{t+1}(x)$ and $P_{(x)}^{(*,0)}$, respectively, and (2.23) can be rewritten as

$$P^{(i)}(x, Y_1(x)) Y_1^{t+1}(x) x + \sum_{i=0}^{i} P^{(i,1)}(x) Y_1^i(x) = 0. \tag{2.24}$$

This equation is similar to (2.18'),but $t \leq s$. Then we can use the same approach to find $y_1, \cdots, y_m$.

For a quasi s-degree equation (2.18'), if y_0 is the s-multiplicity root, we need not solve this equation. Instead, we need to check whether it is a root by Lemma 2.3; otherwise, we need to solve a s-degree equation. Since the s-degree equation has no s-multiplicity root, the multiplicity of any root will be less then s. Thus, we find $y_1, \cdots, y_m$, and the total multiplicity of all roots is equal to or less than s. By repeatedly using Lemma 3.3 to find all $y_1, \cdots, y_m$, we need to solve at most h_i-degree equations for $i = 1, 2, \cdots$ and $i \leq s$, where $2 \leq h_i \leq s$. The worst case is $h_1 = s, \cdots, h_{s-2} = 2$.

3 Divide-and-Conquer Algorithm for the Special Factorization Problem

In Section II, we have derived an algorithm for the special factorization of bivariate polynomial with complexity $O(s^3n^2)$. In this section, using the divide-and-conquer technique and the fast polynomial multiplication algorithm [6, Section 8.3], we improve the algorithm to a divide-and-conquer algorithm for the special factorization of a bivariate polynomial with complexity $O(s^3 n \ln^2 n)$.

Let $Q(x,y) = \sum_{i=0}^{s} P_i(x) y^s$. The special factorization problem is to find all solutions $Y(x) = \sum_{j=0}^{k-1} y_j x^j$ such that

$$\sum_{i=0}^{s} P^i(x) Y^i(x) \equiv 0. \tag{3.1}$$

Since the problem is much simpler than the general factorization problem, it is possible to find an algorithm with a much lower complexity than the general factorization algorithm. In the following, we propose a divide-and-conquer algorithm for the special factorization problem. First, we introduce the algorithm for the case of $s = 1$.

When $s = 1$,

$$Q(x,y) = A(x)y + B(x).$$

Let $A(x) = \sum_{j=0}^{m} a_j x^j$ and $B(x) = \sum_{j=0}^{m+k-1} b_j x^j$. Then, the special factorization problem is to find $Y(x) = \sum_{j=0}^{k-1} y_j x^j$ such that

$$A(x)Y(x) + B(x) = (\sum_{j=0}^{m} a_j x^j)(\sum_{j=0}^{k-1} y_j x^j) + \sum_{j=0}^{m+k-1} b_j x^j \equiv 0. \tag{3.2}$$

In this case, (3.2) has at most one solution. One approach uses the polynomial division algorithm. To divide $B(x)$ by $A(x)$, we can obtain the quotient and the remainder. If the remainder is zero, then the quotient is the solution of (3.2); otherwise, (3.2) has no solution. Using this algorithm, the complexity is $O((m+k)^2)$. We also can use the fast division algorithm to solve this problem; the complexity is $O((m+k)\log^2(m+k))$.

Assume that $k = 2^l$. To find a solution of (3.2), we first find a solution of the following equation:

$$[A(x)Y(x) + B(x)] \quad mod \quad x^k = 0. \tag{3.3}$$

Then, we check whether $Y(x)$ satisfies (3.2). If it satisfies (3.2), then $Y(x)$ is the solution; otherwise, (3.2) has no solution. To find a solution of (3.3), we use a divide-and-conquer algorithm. We denote the running time function by $F(k)$. First, use the divide-and-conquer algorithm to find a solution $Y'(x) = \sum_{j=0}^{k/2} y_i x^j$ such that

$$[A(x)Y'(x) + B(x)] \quad mod \quad x^{k/2} = 0. \tag{3.3'}$$

The running time function is $F(k/2)$. Then, calculate the quotient of $A(x)Y'(x)$ divided by $x^{k/2}$. We denote it by $C(x)$. It is equivalent to calculate a convolution of two polynomials with degree $k/2$. Using the fast convolution algorithm, the complexity is $O(k \log k)$.

Then, use the divide-and-conquer algorithm to find a solution $Y^*(x) = \sum_{j=k/2}^{k} y_i x^{j-k/2}$ such that

$$[A(x)Y^*(x) + B(x)] \quad mod \quad x^{k/2} + C(x) = 0. \tag{3.3''}$$

The running time function is also $F(k/2)$. It can be easily checked that $Y(x) \underline{\underline{\triangle}} Y'(x) + Y^*(x)x^{k/2}$ is the solution of (3.3). Now we are going to calculate $F(k)$. We have

$$F(k) = 2F(k/2) + ck\log k \quad and \quad F(1) = 1.$$

From this recurrence equation, we have $F(k) = O(k\log^2 k)$.

On the other hand, to check whether the obtained $Y(x)$ satisfies (3.2) is equivalent to calculating a convolution of two polynomials with degree $m + k$. Using the convolution algorithm, the complexity is $O((m + k)\log(m + k))$. Therefore, the total complexity is $O((m + k)\log^2(m + k))$.

4 Conclusions

In this paper, we have derived a divide-and-conquer algorithm for a fast factorization of a bivariate polynomial with $y-$degree s. The complexity is $O(s^3 n\log^2 n)$. Considering the application of the generalized Welch-Berlekamp algorithm and fast factorization algorithm to the decoding of Reed-Solomon codes over $GF(2^m)$ beyond the true minimum distance, the total complexity is $O(n\log^3 n)$.

References

[1] V. Guruswami and M. Sudan, "Improved Decoding of Reed-Solomon and Algebraic-Geometric Codes,"Preprint, 1998.

[2] M. Sudan,"Decoding Reed-Solomon Codes beyond the Error-Correction Diameter," *Proc. 35th Annual Allerton Conference on Communication, Control and Computing, 1997* pp. 215-224, 1997.

[3] M. Sudan, "Decoding Reed-Solomon Codes beyond the Error-Correction bound," *Journal of Complexity*, 13(1), pp.180-193, March 1997.

[4] W. S. Feng and R. E. Blahut, "Some Results on the Sudan Algorithm," *Proceedings of 1998 IEEE International Symposium of Information Theory,* MIT, USA, pp.57-57, Aug. 1998.

[5] W. S. Feng and R. E. Blahut, "An Analysis of the Sudan Algorithm," Preprint.

[6] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley Publishing Company, 1976.

Combinatorial Bounds for List Decoding

Venkatesan Guruswami * Johan Håstad † Madhu Sudan ‡ David Zuckerman §

Abstract

Informally, an error-correcting code has "nice" list-decodability properties if every Hamming ball of "large" radius has a "small" number of codewords in it. Here, we report linear codes with non-trivial list-decodability: i.e., codes of large rate that are nicely list-decodable, and codes of large distance that are *not* nicely list-decodable.

1 Introduction

List decoding was introduced independently by Elias [3] and Wozencraft [15] as a relaxation of the "classical" notion of decoding by allowing a list of codewords as answers. The decoding is considered successful as long as the correct message is included in the list. Formally, an instance of the list decoding problem for a code **C** consists of a received word **r** and an error bound e, and the goal is to output a list of *all* codewords of **C** that are within a Hamming distance of e from **r**.

Early works in this area [3, 15, 10] focused on the probability of error in this more general decoding model. Research in the eighties applied this notion in a more adversarial setting and investigated what happens if the error is effected by an adversary or a "jammer", as opposed to a probabilistic channel. Works of Zyablov and Pinsker [16], Blinovskii [1] and Elias [4] applied in this setting. (The paper by Elias [4] also gives a very good summary of the prior work and history.) The basic (purely combinatorial) question raised in this setting is: How many errors could still be recovered from, with lists of small size? Two basic parameters of interest thus are the number of errors and the allowed size of the output list. These parameters are usually studied as a function of some of the more classical parameters of error-correcting codes. How large can the rate of a code be if we want small list sizes for a certain number of errors? And how do codes of large minimum distance perform with respect to list decoding? Recently there has been rejuvenated interest in this line of work thanks to the development of some efficient algorithms for list decoding in [12, 6, 11, 7]. These algorithms decode with polynomial sized lists (and sometimes with constant sized lists) for a number of errors that is much more than half the minimum distance of the code, and investigations of the tightness of the algorithms have led Hoholdt and Justesen [9] to re-initiate the investigation of the combinatorial bounds on list decoding.

In this paper we continue the investigation of bounds on list decoding. In particular, we investigate codes that exhibit non-trivial list decoding performance. Specifically, we report the existence of linear codes

---

[*] Laboratory for Computer Science, 200 Technology Square, Cambridge, MA 02139, USA. Email: `venkat@theory.lcs.mit.edu`.

[†] Department of Numerical Analysis and Computer Science, Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Email: `johanh@nada.kth.se`. Supported in part by the Göran Gustafsson foundation.

[‡] Laboratory for Computer Science, 200 Technology Square, Cambridge, MA 02139, USA. Email: `madhu@mit.edu`. Supported in part by a Sloan Foundation Fellowship and NSF Career Award CCR-9875511.

[§] Department of Computer Science, University of Texas, Austin, TX 78712, USA. Email: `diz@cs.utexas.edu`. Currently on leave at the University of California, Berkeley. Supported in part by NSF Grant CCR-9912428, NSF NYI Grant CCR-9457799, and a David and Lucile Packard Fellowship for Science and Engineering.

of large rate that are nicely list-decodable, and codes of large minimum distance which are not nicely list-decodable (the precise quantitative versions of these results are stated in the next section). To motivate this study we first fix some standard notation and then define two fundamental questions (parameters) to study in the context of list decoding.

2 Definitions and Main Results

For a prime power q, let $\mathbb{F}_q$ denote a finite field of cardinality q. An $[n,k]_q$ (linear) code C is a k-dimensional vector space in $\mathbb{F}_2^n$. Unless explicitly mentioned otherwise, throughout this paper, we will only be interested in linear codes and will moreover restrict ourselves to the binary case (when $q = 2$).

For vectors $x, y \in \mathbb{F}_2^n$, let $\text{distance}(x,y)$ denote the Hamming distance between them. The minimum distance of a code C, denoted $\text{distance}(C)$, is the quantity $\min_{x,y \in C, x \neq y}\{\text{distance}(x,y)\}$.

Since the main thrust of this paper is the asymptotic performance of the codes, we define analogs of the quantities above for infinite families of codes. An infinite family of (binary) codes is a family $\mathcal{C} = \{C_i | i \in \mathbb{Z}^+\}$ where C_i is an $[n_i, k_i]_2$ code with $n_i > n_{i-1}$. We define the *rate* of an infinite family of codes $\mathcal{C}$ to be $\text{rate}(\mathcal{C}) = \inf_i \left\{\frac{k_i}{n_i}\right\}$, and the *(relative) distance* of an infinite family of codes $\mathcal{C}$ to be $\Delta(\mathcal{C}) = \inf_i \left\{\frac{\text{distance}(C_i)}{n_i}\right\}$.

We now define the list decoding radius of a code. For non-negative integer r and $x \in \mathbb{F}_2^n$, let $B(x,r)$ denote the ball of radius r around x, i.e., $B(x,r) = \{y \in \mathbb{F}_2^n | \text{distance}(x,y) \leq r\}$.

Definition 1 (List decoding Radius) *For an* $[n,k]$ *binary code* C*, and list size* ℓ*, the* list of ℓ decoding radius *of* C*, denoted* $\text{radius}(C,\ell)$ *is defined to be*

$$\text{radius}(C,\ell) = \max\left\{e | \forall x \in \mathbb{F}_2^n,\ |B(x,e) \cap C| \leq \ell\right\}.$$

For a function $\ell : \mathbb{Z}^+ \to \mathbb{Z}^+$*, and an infinite family of codes* $\mathcal{C}$*, let the list of* ℓ *decoding radius of* $\mathcal{C}$*, denoted* $\text{Rad}(\mathcal{C},\ell)$ *be the quantity*

$$\text{Rad}(\mathcal{C},\ell) = \inf_i \left\{\frac{\text{radius}(C_i, \ell(n_i))}{n_i}\right\}.$$

It is interesting to study the list decoding radius of infinite families of codes as a function of their distance and rate, when the list size is either bounded by a constant, or grows as polynomially in the length of the code. Within this scope the broad nature of the two main questions are: (1) Do there exist codes of large rate with large list decoding radius for a fixed function ℓ? and (2) Do there exist codes of large distance with small list decoding radius for a given function ℓ? Note that the other two questions are uninteresting: specifically, it is possible to construct codes of small rate that have small list decoding radius (for example, by taking a linear code and just adding few codewords at a small distance to some existing codeword); and it is possible to construct codes of small distance that have large list decoding radius even for lists of size 2 (for example by taking a code of large minimum distance and adding one codeword at a small distance to some existing codeword). In what follows we introduce some formal parameters to study the above questions.

2.1 List decoding radius vs. Rate of the code

Definition 2 (Upper bound on list decoding radius) *For real rate* $0 \leq R \leq 1$ *and list size* $l : \mathbb{Z}^+ \to \mathbb{Z}^+$*, the* upper bound on list of ℓ decoding radius *for (binary) codes of rate* R*, denoted* $U_\ell(R)$*, is defined to be*

$$U_\ell(R) = \sup_{\mathcal{C} \,|\, \text{rate}(\mathcal{C}) \geq R} \text{Rad}(\mathcal{C},\ell).$$

Note that the reason for the term "upper bound" is that $U_\ell(R)$ is the list decoding radius of the *best code* (i.e. one with largest possible list decoding radius) among codes that have at least a certain rate. The case where the list size function is a constant or a polynomial is of special interest to us, and we consider the following definitions.

Definition 3 *For real rate* $0 \le R \le 1$ *and constant c, the quantity* $U_c^{\text{const}}(R)$ *is defined to be* $U_\ell(R)$ *where* $\ell(n) = c$. *The quantity* $U_c^{\text{poly}}(R)$ *is defined to be* $\sup_{c_1} U_{\ell_{c_1}}(R)$ *where* $\ell_{c_1}(n) = c_1 n^c$. *The quantity* $U^{\text{const}}(R)$ *(resp.* $U^{\text{poly}}(R)$*) will denote the quantity* $\limsup_{c\to\infty}\{U_c^{\text{const}}(R)\}$ *(resp.* $\limsup_{c\to\infty}\{U_c^{\text{poly}}(R)\}$*).*

Thus the quantities $U^{\text{const}}(R)$ and $U^{\text{poly}}(R)$ denote the maximum possible value of the (relative) list decoding radius for lists of constant and polynomial size, respectively. These quantities are actually surprisingly well-understood. Zyablov and Pinsker [16] showed that $U^{\text{const}}(R) = U^{\text{poly}}(R) = H^{-1}(1-R)$ which matches the Gilbert-Varshamov bound (here $H(\cdot)$ is the binary entropy function: $H(x) = -x\lg x - (1-x)\lg(1-x)$). The behavior of the upper bound on list decoding radius for lists of size c, for specific constants c, however, was not known completely. The results of Zyablov and Pinsker [16], stated in our notation, showed that

$$U_c^{\text{const}}(R) \ge H^{-1}\left(1 - \frac{1}{\lg(c+1)} - R\right), \tag{1}$$

(this result clearly implies the above-mentioned result $U^{\text{const}}(R) = H^{-1}(1-R)$) and Elias [4] had shown (again rephrased in our notation) that

$$U_c^{\text{const}}(R) \ge \frac{1}{2}\Big(1 - \sqrt{1 - \frac{2(c-1)}{c}H^{-1}(1-R)}\Big)\,. \tag{2}$$

Here we improve the bounds significantly and show the following. In particular this answers the main open question posed by Elias [4] on whether the bound (1) can be improved.

Theorem 4 *For each fixed integer* $c \ge 1$, *and rate* $0 < R < 1$, $U_c^{\text{const}}(R) \ge H^{-1}(1 - \frac{1}{c} - R)$.

We prove this theorem using the probabilistic/greedy method. We stress that our result is only existential and not constructive. Surprisingly our proof does not give a high-probability result; and it is open whether a random code satisfies the above trade-off between rate and list decoding radius with high probability (the result of Equation (1) above is a high-probability result). For general non-linear codes, the bound of Theorem 4 is shown in [4] and in fact holds with high probability for a random code.

All the above results (including ours from Theorem 4) provide lower bounds on $U_c^{\text{const}}(\cdot)$ (except for the simple upper bound $U_c^{\text{const}}(R) \le U^{\text{poly}}(R) \le H^{-1}(1-R)$). The results of Blinovskii [1] imply non-trivial upper bounds on $U_c^{\text{const}}(\cdot)$ for fixed constants c. His results are not easily stated, but they do imply that $U_c^{\text{const}}(R)$ is *strictly less than* $H^{-1}(1-R)$ for every c, R, and thus in order to achieve the Gilbert-Varshamov bound, we *have to* allow an unbounded number of elements in the list.

Theorem 5 [Follows from [1]] *For every* $c \ge 1$ *and* $0 < R < 1$, *we have* $U_c^{\text{const}}(R) < H^{-1}(1-R)$.

Comment on Theorem 4: Suppose we want a binary code family $\mathcal{C}$ with $\text{Rad}(\mathcal{C}, \ell) = (1/2 - \varepsilon)$ for some constant $\varepsilon > 0$ and where ℓ is the constant function $\ell(n) = c\ \forall n$. Then Theorem 4 tells us that such code families with rate $\Omega(\varepsilon^2)$ exist for a list size of $c = O(\varepsilon^{-2})$. We remark that the result of Elias (Equation 2) would give code families of rate $\Omega(\varepsilon^4)$ for a list size $c = O(\varepsilon^{-2})$, while that of Zyablov and Pinsker (Equation 1) would give code families of rate $\Omega(\varepsilon^2)$ but with a much worse list size of $c = 2^{O(\varepsilon^{-2})}$. One can also show using the results of Blinovskii [1] that in order to have such families $\mathcal{C}$ with $\text{rate}(\mathcal{C}) > 0$, we must have $c = \Omega(\varepsilon^{-2})$. In this sense, the result of Theorem 4 is asymptotically the best possible (for example, the $1/c$ "loss" term cannot be improved to $1/c^{1+\gamma}$ for any positive γ).

2.2 List decoding radius vs. Distance of the code

Next we move on to lower bounds on the list decoding radius. As mentioned earlier, it makes sense to study this as a function of the minimum distance of the code. A large minimum distance implies a large list decoding radius by existing combinatorial bounds (see for example [5]), and we want to find the smallest possible list decoding radius for a code of (at least) a certain minimum distance. This motivates the next definition.

Definition 6 (Lower bound on list decoding radius) *For a distance* $0 \le \delta \le 1$, *and list size* $\ell : \mathbb{Z}^+ \to \mathbb{Z}^+$, *the* lower bound on list of ℓ decoding radius *for (binary) codes of distance* δ, *denoted* $L_{\ell,q}(\delta)$, *is defined to be*

$$L_\ell(\delta) = \inf_{C \mid \Delta(C) \ge \delta} \mathrm{Rad}(C, \ell).$$

Note that both in the case of the upper bound function U_ℓ and the lower bound function L_ℓ one could allow the arguments, i.e., rate and distance to be functions of n, in which case the supremum would be taken over codes C that satisfy $\mathrm{rate}(C_i) \ge R(n_i) \cdot n_i$ (or in the case of the lower bound function, we would take the infimum over codes that satisfy $\Delta(C_i) \ge \delta(n_i) \cdot n_i$).

As in the case of the upper bound function, we introduce notation to study the special cases when the list size is a constant or grows as a polynomial.

Definition 7 *For real distance* $0 \le \delta \le 1/2$ *and constant* c, *the quantity* $L_c^{\mathrm{const}}(\delta)$ *is defined to be* $L_\ell(\delta)$ *where* $\ell(n) = c$. *The quantity* $L_c^{\mathrm{poly}}(\delta)$ *is defined to be* $\sup_{c_1} L_{\ell_{c_1}}(\delta)$ *where* $\ell_{c_1}(n) = c_1 n^c$. *The quantity* $L^{\mathrm{const}}(\delta)$ *(resp.* $L^{\mathrm{poly}}(\delta)$*) will denote the quantity* $\limsup_{c\to\infty}\{L_c^{\mathrm{const}}(\delta)\}$ *(resp.* $\limsup_{c\to\infty}\{L_c^{\mathrm{poly}}(\delta)\}$*).*

Note that we restrict $\delta \le 1/2$ since binary codes with minimum distance $\delta > 1/2$ have only a polynomial number of codewords and are thus not interesting. It is clear that $L_1(\delta) = \delta/2$. It is also easy to see that $L^{\mathrm{poly}}(\delta) \le \delta$ (since there exist codes of minimum distance δ with super-polynomially many codewords at minimum distance.) Thus all lower bounds of interest lie in the range $[\delta/2, \delta]$. The exact values are, however, mostly unknown. The main motivation for our work is the following conjecture.

Conjecture 8 *For every* $0 < \delta < 1/2$, $L^{\mathrm{const}}(\delta) = L^{\mathrm{poly}}(\delta) = \frac{1}{2} \cdot (1 - \sqrt{1 - 2\delta})$.

Evidence in support of the conjecture comes piecemeal. Firstly, it is known that $L^{\mathrm{poly}}(\delta) \ge L^{\mathrm{const}}(\delta) \ge L_2^{\mathrm{const}}(\delta) \ge \frac{1}{2} \cdot \left(1 - \sqrt{1 - 2\delta}\right)$ (see for example [5]). Upper bounds on L^{poly} and L^{const} are not as well studied. Justesen and Hoholdt [9] demonstrate some MDS code families C of distance δ with $\mathrm{Rad}(C, c) \le (1 - \sqrt{1 - \delta})$, but this does not apply for codes over any fixed size alphabet, and in particular for binary codes.

The quantity $L^{\mathrm{poly}}(\delta)$ is even less well understood. When δ is either very large (of the form $1/2 - o(1)$) or very small (of the form $o(1)$), there is some evidence confirming this bound. In particular, Dumer et al. [2] construct a family of linear codes C, for any $\varepsilon > 0$, for which $\delta(n) = n^{\varepsilon - 1}$ and $L^{\mathrm{poly}}(\delta) \le \delta/(2 - \varepsilon)$ which matches the conjecture above reasonably closely. In unpublished work, Ta-Shma and Zuckerman [13] show that for every $\varepsilon > 0$, $L^{\mathrm{poly}}(\frac{1}{2}(1 - n^{\varepsilon - 1/2})) \le \frac{1}{2}(1 - \frac{1}{3\varepsilon} n^{\varepsilon - 1/2})$. This seems to show that the tangent of the curve $L^{\mathrm{poly}}(\delta)$ has infinite slope as $\delta \to 1/2$, which is consistent with the conjecture above (and thus mild evidence in favor of the conjecture). One additional reason for believing in the conjecture is that if the definition of codes is extended to allow non-linear codes, then indeed it is known that the conjecture is true (see for example [5]). All this evidence adds support to the conjecture, however remains far from proving it. In fact till this paper it was not even known if $L_c^{\mathrm{poly}}(\delta) < \delta$. The following theorem resolves this question.

Theorem 9 *For every* c *and every* δ, *we have* $L_c^{\mathrm{poly}}(\delta) < \delta$.

Further, for the case $\delta = \frac{1}{2} \cdot (1 - o(1))$, we actually get close to proving the above conjecture. The theorem below follows from Lemma 12 which is stated and proved in Section 3.3.

Theorem 10 *For $\delta = \frac{1}{2}(1 - o(1))$ for some (explicitly given) $o(1)$ function, for every $\varepsilon > 0$, we have $L^{\rm poly}(\delta) \le \frac{1}{2}[1 - (1 - 2\delta)^{1/2+\varepsilon}]$.*

Organization of the Paper. We study the lower bound functions $L^{\rm poly}(\delta)$ and $L_c^{\rm poly}(\delta)$ in Section 3 and prove Theorems 9 and 10. In Section 4, we study the function $U_c^{\rm const}(R)$ and prove Theorem 4.

3 List Decoding Radius and Minimum Distance

We now prove upper bounds on the function $L^{\rm poly}(\delta)$ claimed in Theorems 9 and 10. We will first prove Theorem 10 which shows that when $\delta = \frac{1}{2} \cdot (1 - o(1))$, one "almost" has a proof of Conjecture 8. A modification of this proof will also yield the proof of Theorem 9. We first review the basic definitions and concepts from (Discrete) Fourier analysis that will be used in our proof.

3.1 Fourier analysis and Group characters

For this section, it will be convenient to represent Boolean values by $\{1, -1\}$ with 1 standing for FALSE and -1 for TRUE. This has the nice feature that XOR just becomes multiplication. Thus a binary code of blocklength m will be a subset of $\{1, -1\}^m$. There are 2^t linear functions $\chi_\alpha : \{0,1\}^t \to \{1, -1\}$ on t-variables, one for each $\alpha \in \{0,1\}^t$. The function χ_α is defined by $\chi_\alpha(x) = (-1)^{\alpha \cdot x} = (-1)^{\sum \alpha_i x_i}$. Fixing some representation of the field $\mathrm{GF}(2^t)$ as elements of $\{0,1\}^t$, the linear functions χ_α are the *additive characters* of the field $\mathrm{GF}(2^t)$, and can also be indexed by elements $\alpha \in \mathrm{GF}(2^t)$. We will do so in the rest of the paper. We also have, for each $y \in \mathrm{GF}(2^t)$, $\sum_\alpha \chi_\alpha(y)$ equals 0 if $y \neq 0$ and 2^t if $y = 0$, where the summation is over all $\alpha \in \mathrm{GF}(2^t)$.

We can define an inner product $\langle f, g \rangle$ for functions $f, g : \mathrm{GF}(2^t) \to \mathbb{R}$ as $\langle f, g \rangle = 2^{-t} \sum_x f(x) g(x)$. The linear functions form an orthonormal basis for the space of real-valued functions on $\mathrm{GF}(2^t)$ with respect to this inner product. Thus every real-valued function on $\mathrm{GF}(2^t)$, and in particular every Boolean function $f : \mathrm{GF}(2^t) \to \{1, -1\}$ can be written in terms of the χ_α's as: $f(x) = \sum_{\alpha \in \mathrm{GF}(2^t)} \hat{f}_\alpha \chi_\alpha(x)$. The coefficient f_α is called the *Fourier coefficient* of f with respect to α and satisfies $\hat{f}_\alpha = \langle f, \chi_\alpha \rangle = 2^{-t} \sum_x f(x) \chi_\alpha(x)$. If we define the distance between functions f, g as $\Delta(f, g) = \Pr_x [f(x) \neq g(x)]$, then $\hat{f}_\alpha = 1 - 2\Delta(f, \chi_\alpha)$. The Fourier coefficients of a Boolean function also satisfy the Plancherel's identity $\sum_\alpha \hat{f}_\alpha^2 = 1$.

Hadamard code: For any integer t, the Hadamard code Had_t of dimension t maps t bits (or equivalently elements of $\mathrm{GF}(2^t)$) into $\{1, -1\}^{2^t}$ as follows: For any $x \in \mathrm{GF}(2^t)$, $\mathrm{Had}_t(x) = \langle \chi_\alpha(x) \rangle_{\alpha \in \mathrm{GF}(2^t)}$.

3.2 Idea behind the Construction

Since our aim is to prove lower bounds on the list decoding radius we must construct codes with large minimum distance with a large number of codewords in a ball of desired radius. The specific codes we construct are obtained by concatenating an outer Reed-Solomon code over a finite field $F = \mathrm{GF}(2^t)$ with the Hadamard code Had_t of blocklength 2^t and dimension t. Thus the messages of this code will be degree ℓ polynomials over $\mathrm{GF}(2^t)$ for some ℓ, and such a polynomial P is mapped into the codeword $\langle \mathrm{Had}_t(P(z_1)), \ldots, \mathrm{Had}_t(P(z_{2^t})) \rangle$ where $z_1, z_2, \ldots, z_{2^t}$ is some enumeration of the elements in $\mathrm{GF}(2^t)$.

It is easy to see that this code has blocklength 2^{2t} and minimum distance $\frac{1}{2}(1 - \frac{\ell}{n})2^{2t}$. If $\ell = (1 - 2\delta)n$, then the relative minimum distance is δ, and for future reference we denote this code by RS-HAD$_t(\delta)$.

To construct the *received word* (which will be the center of the Hamming ball with a lot of codewords), consider the following. Suppose we could pick an appropriate subset S of $\mathrm{GF}(2^t)$ and construct a Boolean function $f : \mathrm{GF}(2^t) \to \{1,-1\}$ that has large Fourier coefficient $\hat{f}_\alpha$ with respect to α for $\alpha \in S$. Let $\mathbf{v} \in \{1,-1\}^{2^t}$ be the 2^t-dimensional vector consisting of the values of f on $\mathrm{GF}(2^t)$. The word $\mathbf{v}^{|F|}$, i.e., $\mathbf{v}$ repeated $|F|$ times will be the "received word" (the center of the Hamming ball which we want to show has several codewords). Since f has large Fourier support on S, $\mathbf{v}^{|F|}$ will have good agreement with all codewords that correspond to messages (polynomials) P that satisfy $P(z_i) \in S$ for many field elements z_i. By picking for the set S a multiplicative subgroup of $\mathrm{GF}(2^t)$ suitable size, we can ensure that there are several such polynomials, and hence several codewords in the concatenated code with good agreement with $\mathbf{v}^{|F|}$.

The main technical component of our construction and analysis is the following Theorem which asserts the existence of Boolean functions f with large support on subgroups S of $\mathrm{GF}(2^t)$. We will defer the proof of the theorem to Section 3.5, and first use it to prove Theorems 10 and 9.

Theorem 11 *There exist infinitely many integers s with the following property: For infinitely many integers t, there exists a multiplicative subgroup S of $\mathrm{GF}(2^t)$ of size s such that the following holds: For every $\beta \neq 0$ in $\mathrm{GF}(2^t)$ there exists a function $f : \mathrm{GF}(2^t) \to \{1,-1\}$ with $\sum_{\alpha \in \beta S} \hat{f}_\alpha \geq \sqrt{\frac{s}{3}}$. (Here βS denotes the coset $\{\beta x : x \in S\}$ of S.)*

Remarks: Our proof of the above theorem in fact gives the following additional features which we make use of in our applications of the theorem.

1. The integers s exists with good density; in particular for any integer $k \geq 4$, there exists an s, with $k \leq s < 3k$, that satisfies the requirements of Theorem 11.
2. We can also add the condition that there exist infinitely many t including one that is at most s, and the theorem still holds.

For any subset $S \subseteq \mathrm{GF}(2^t)$, one can show that $\sum_{\alpha \in S} \hat{f}_\alpha$ is at most $|S|^{1/2}$ using Plancherel's identity and Cauchy-Schwartz, and Theorem 11 shows that we can achieve a sum of $\Omega(|S|^{1/2})$ infinitely often for appropriate multiplicative subgroups S.

3.3 Proof of Theorem 10

We now employ Theorem 11 to prove Theorem 10. We in fact prove the following Lemma which clearly establishes Theorem 10.

Lemma 12 *For every $\varepsilon > 0$, there exist infinitely many t such that the following holds: Let $N = 2^{2t}$. There exists a vector $\mathbf{r} \in \{1,-1\}^N$ and $\delta = \frac{1}{2}(1 - (\log N)^{-e})$ for some $e < 1$, such that the number of codewords C of the code* RS-HAD$_t(\delta)$ *with* distance$(\mathbf{r}, C) \leq \frac{N}{2}(1 - (1-2\delta)^{1/2+\varepsilon})$ *is at least $N^{\Omega(\log^{\varepsilon} N)}$.*

Proof: Let s, t be any pair of integers guaranteed by Theorem 11 with $t \leq s$ (we are using one of the remarks following Theorem 11 here). Let S be a multiplicative subgroup of $\mathrm{GF}(2^t)$ of size s and $f : \mathrm{GF}(2^t) \to \{1,-1\}$ a function such that

$$\sum_{\alpha \in S} \hat{f}_\alpha \geq \sqrt{\frac{s}{3}}\,. \tag{3}$$

Let $n = 2^t$, $N = 2^{2t}$ and $p = (n-1)/s$. Then $S \cup \{0\}$ consists of all elements in $\mathrm{GF}(2^t)$ which are p'th powers of some element of $\mathrm{GF}(2^t)$.

We first fix the "received word" $\mathbf{r}$. Let $\mathbf{v} \in \{1,-1\}^n$ be the vector $\langle f(x) \rangle_{x \in \mathrm{GF}(2^t)}$ of all values of f. Then $\mathbf{r} = \mathbf{v}^n$, i.e. the vector $\mathbf{v}$ repeated $n = 2^t$ times, one for each position of the outer Reed-Solomon code.

Let δ be a parameter to be specified later and $\ell = (1-2\delta)n$. Consider the binary code $\mathbf{C} = \text{RS-HAD}_t(\delta)$ obtained by concatenating a Reed-Solomon code of dimension $\ell + 1 = (1 - 2\delta)n + 1$ over $\mathrm{GF}(2^t)$ with Had_t. $\mathbf{C}$ has blocklength N and minimum distance δN. We now want to demonstrate several codewords in $\mathbf{C}$ that are "close" to $\mathbf{r}$. We prove this picking codewords in $\mathbf{C}$ at random from some distribution and showing that the agreement with $\mathbf{r}$ is "large" with good probability.

Let $m = \lfloor \ell/p \rfloor$ and consider a message (degree ℓ polynomial over $\mathrm{GF}(2^t)$) P of $\mathbf{C}$ which is of the form $P(x) = R(x)^p$ for a *random* polynomial R of degree at most m over $\mathrm{GF}(2^t)$. The Reed-Solomon encoding $(b_1, b_2, \ldots, b_n)$ of P satisfies $b_i \in S \cup \{0\}$ for every i, $1 \le i \le n$. Moreover, for each i and each $a \in S$, we have $\mathbf{Pr}[b_i = a] = p/n$, and $\mathbf{Pr}[b_i = 0] = 1/n$. Further, the choices of b_i are *pairwise independent*. Now, using Equation (3) and a simple probabilistic analysis one can show that, for large enough N, with probability at least $1/2$ (over the choice of the polynomial R), the encoding of P differs from $\mathbf{r}$ in at most $(\frac{1}{2} - \frac{1}{2\sqrt{4s}})N$ codeword positions.

We thus get at least $\frac{1}{2}n^m$ codewords in a ball of radius $(\frac{1}{2} - \frac{1}{2\sqrt{4s}})N$. We now pick parameters suitably to conclude the result. We have $t = \frac{1}{2}\log N$ and $s \ge t$. Picking $m \le s^{\varepsilon}$, we have $(1 - 2\delta) \simeq m/s = s^{\varepsilon - 1} \ge \log^e N$ for some constant $e < 1$, and we can thus have a minimum distance of $\delta = \frac{1}{2}(1 - (\log N)^{-e})$ for some $e < 1$. Also for ε small enough, $(1 - 2\delta)^{1/2+\varepsilon} \simeq s^{(\varepsilon-1)(1/2+\varepsilon)} \le (4s)^{-1/2}$ and thus we have $\Omega(n^m) = N^{\Omega(\log^{\varepsilon} N)}$ codewords in a Hamming ball of radius $\frac{N}{2}(1 - (1 - 2\delta)^{1/2+\varepsilon})$. $\square$

3.4 Proof of Theorem 9

We now turn to obtaining upper bounds on $L_c^{\mathrm{poly}}(\delta)$ for a fixed constant c. One way to achieve this would be to pick $m \simeq 2c$ in the above proof, and then pick $s \simeq 2c/(1-2\delta)$ and this would give (roughly) $L_c^{\mathrm{poly}}(\delta) \le \frac{1}{2}(1 - \left(\frac{1-2\delta}{6c}\right)^{1/2})$. However this upper bound is better than δ only for δ large enough, specifically for $\delta > \frac{1}{2} - \frac{1}{12c}$. We thus have to modify the construction of Lemma 12 in order to prove Theorem 9. We prove the following lemma which will in turn imply Theorem 9. Since our goal was only to establish Theorem 9, we have not attempted to optimize the exact bounds in the lemma below.

Lemma 13 *For every c and every δ, we have*

$$L_c^{\mathrm{poly}}(\delta) \le \min_{0 \le \alpha \le 1/2-\delta} \Big\{(\delta + \alpha)(1 - (\frac{\alpha}{12(2c+1)})^{1/2})\Big\}.$$

Proof of Lemma 13: To prove the claimed upper bound on $L_c^{\mathrm{poly}}(\delta)$, we will closely follow the construction from the proof of Lemma 12. Let $0 < \delta < 1/2$, $0 \le \alpha \le (1/2 - \delta)$, and c be given. Define $\alpha' = 2\alpha$ and pick an integer s, $2(2c+1)/\alpha' \le s < 6(2c+1)/\alpha'$ such that the conditions of Theorem 11 are met (we know such an s exists by the remarks following Theorem 11). Let t be *any* integer for which a subgroup S of $\mathrm{GF}(2^t)$ exists as guaranteed by Theorem 11 (there are once again infinitely many such values of t).

Now we describe the actual construction for a particular δ, α', s, t. Let $n = 2^t$, $N = n^2$ and $p = (n-1)/s$. Let S be a multiplicative subgroup of $\mathrm{GF}(2^t)$ of size s. As in the proof of Lemma 12, the code will again be $\text{RS-HAD}_t(\delta)$ (the messages of the code will thus be polynomials over $\mathrm{GF}(2^t)$ of degree at most $\ell = (1 - 2\delta)n$ and the code has blocklength N). The only change will be in the construction of the received word $\mathbf{r}$. Let $B = (1 - 2\delta - \alpha')n$. We set the first B blocks of $\mathbf{r}$ to be all zeroes. The last $(n - B)$ blocks of $\mathbf{r}$ will be vectors $\mathbf{v}^{(i)} \in \{1, -1\}^{2^t}$, $B < i \le n$, defined as follows. Let $z_1, z_2, \ldots, z_B$ be the B elements of $\mathrm{GF}(2^t)$ that correspond to the first B positions of the Reed-Solomon code. For $B < i \le n$, define $\beta_i = (z_i - z_1)\cdots(z_i - z_B)$, and $f^{(i)} : \mathrm{GF}(2^t) \to \{1, -1\}$ be a function with large Fourier support on the coset $\beta_i S$ of S as guaranteed by Theorem 11; i.e. $\sum_{\alpha \in \beta_i S} \hat{f}_{\alpha}^{(i)} \ge \sqrt{s/3}$. We set $\mathbf{v}^{(i)}$ to be the table of values of $f^{(i)}$.

Let $m = 2c + 1$. We will consider the messages corresponding to polynomials of the form $P(x) = (x - z_1) \cdots (x - z_B) R(x)^p$ where $z_1, \ldots, z_B$ of $\mathrm{GF}(2^t)$ are the B elements of $\mathrm{GF}(n)$ that correspond to the first B positions of the Reed-Solomon code and R is a random degree m polynomial. Note that $\mathrm{degree}(P) = B + pm = \ell - \alpha' n + \frac{n-1}{s}(2c+1) \leq \ell$ since we picked $s \geq 2(2c+1)/\alpha'$. Note that the encoding of every such P agrees with $\mathbf{r}$ in the first B blocks.

Using arguments similar to those in the proof of Lemma 12, one can show that with high probability (say at least 1/2), the codeword P differs from $\mathbf{r}$ in at most $E = (n - B)(\frac{1}{2} - \frac{1}{2\sqrt{4s}})n$ positions, and thus there are at least $\frac{1}{2}n^m$ codewords within a ball of radius E around $\mathbf{r}$. Since $N = n^2$, $m = 2c + 1$ and $s < 6(2c+1)/\alpha'$, we have $\omega(N^c)$ codewords in a Hamming ball of radius $N(\delta + \alpha'/2)(1 - \sqrt{\frac{\alpha'}{24(2c+1)}})$, and recalling that $\alpha' = 2\alpha$, the claimed result follows. □ *(Lemma 13)*

Proof of Theorem 9: We want to prove $L_c^{\mathrm{poly}}(\delta) < \delta$. Note that when $\delta > \frac{1}{2} - \frac{1}{48(2c+1)}$, setting $\alpha = 1/2 - \delta$ gives

$$L_c^{\mathrm{poly}}(\delta) \leq \frac{1}{2}\left(1 - \left(\frac{1-2\delta}{24(2c+1)}\right)^{1/2}\right) < \delta\,.$$

When $\delta \leq \frac{1}{2} - \frac{1}{48(2c+1)}$, setting $\alpha = \delta^2/48(2c+1)$ (this is a valid setting since it is less than $1/2 - \delta$), we have $L_c^{\mathrm{poly}}(\delta) \leq \delta + \alpha - \delta(\frac{\alpha}{12(2c+1)})^{1/2} < \delta$. Thus we have $L_c^{\mathrm{poly}}(\delta) < \delta$ in either case. □ *(Theorem 9)*

3.5 Proof of Theorem 11

The proof proceeds in several steps. We first state the following lemma which shows that if a subset S of $\mathrm{GF}(2^t)$ satisfies a certain property, then there exists a Boolean function $f : \mathrm{GF}(2^t) \to \{1, -1\}$ such that $\sum \hat{f}_\alpha$ is large when summed over $\alpha \in S$. The proof of the lemma is omitted here for reasons of space and can be found in the full version of the paper.

Lemma 14 *For any integer t, let S be any subset of elements of the field* $\mathrm{GF}(2^t)$ *such that no four (distinct) elements of S sum up to 0. Then there exists a function $f : \mathrm{GF}(2^t) \to \{1, -1\}$ with $\sum_{\alpha \in S} \hat{f}_\alpha \geq \sqrt{|S|/3}$.*

Given the statement of Lemma 14, we next turn to constructing subgroups of $\mathrm{GF}(2^t)$ with the property that no four (or fewer) distinct elements of the subgroup sum up to 0. To construct such subgroups, we make use of the following simple lemma about the existence of certain kinds of cyclic codes.

Lemma 15 *Let $k \geq 4$ be any integer. Then there exists an integer s in the interval $[k, 3k)$ such that a maximal binary BCH code of blocklength s and minimum distance at least 5 exists.*

Proof: Let s be an integer of the form $2^f - 3$ in the range $[k, 3k)$ (such an integer clearly exists). Let β be the primitive s'th root of unity over $\mathrm{GF}(2)$ and let h be the minimal polynomial of β over $\mathrm{GF}(2)$. Clearly, $h(\beta^{2^i}) = 0$ for all $i \geq 1$, and hence $h(\beta^2) = h(\beta^4) = 0$. Since $\beta^{2^f} = \beta^3$, we also have $h(\beta^3) = 0$. Now the consider the cyclic code C_h of blocklength s with generator polynomial h. It is clearly maximal since h, being the minimal polynomial of β, is irreducible over $\mathrm{GF}(2)$. Also $h(\beta^i) = 0$ for $i = 1, 2, 3, 4$. Using the BCH bound on designed distance (see, for example, Section 6.6 of [14]), this implies that the minimum distance of C_h is at least 5, as desired. □

Lemma 16 *Let $k \geq 4$ be any integer. Then there exists an integer s in the interval $[k, 3k)$ with the following property. For infinitely many integers t, including some integer which is at most s, there exists a multiplicative subgroup S of* $\mathrm{GF}(2^t)$ *of size s such that no four or fewer distinct elements of S sum up to 0 (in* $\mathrm{GF}(2^t)$*). Moreover, for any non-zero $\beta \in \mathrm{GF}(2^t)$ this property holds for the coset βS as well.*

Proof: Given k, let $k \leq s < 3k$ be an integer for which there exists a binary BCH code C of blocklength s as guaranteed by Lemma 15 exists. Such a code is generated by an irreducible polynomial h where $h(x)|(x^s - 1)$. Let $t = \text{degree}(h)$; clearly $t \leq s$. Consider the finite field $F = \mathbb{F}_2[X]/(h(X))$ which is isomorphic to $\text{GF}(2^t)$, and consider the subgroup S of size s of F comprising of $\{1, X, X^2, X^3, \ldots, X^{s-1}\}$. The fact that C has distance at least 5 implies that $\sum_{i \in G} X^i$ is not divisible by $h(X)$ for any set G of size at most 4, and thus no four or fewer distinct elements of S sum up to 0 in the field F. This gives us one value of $t \leq s$ for which the conditions of Lemma 16 are met, but it is easy to see that any multiple of t also works, since the same S is also a (multiplicative) subgroup of $\text{GF}(2^{kt})$ for all $k \geq 1$. The claim about the cosets also follows easily, since if $a_1 + a_2 + a_3 + a_4 = 0$ where each $a_i \in \beta S$, then $\beta^{-1}a_1 + \beta^{-1}a_2 + \beta^{-1}a_3 + \beta^{-1}a_4 = 0$ as well, and since $\beta^{-1}a_i \in S$, this contradicts the property of S. □

We now have all the ingredients necessary to easily deduce Theorem 11.

Proof of Theorem 11: Theorem 11 now follows from Lemma 14 and Lemma 16. Note also that the statement of Lemma 16 implies the remarks made after the statement of Theorem 11. □ *(Theorem 11)*

4 List Decoding Radius vs. Rate

We now prove Theorem 4.

Proof of Theorem 4: For each fixed integer $c \geq 1$ and $0 < p < 1/2$, we use the probabilistic method to guarantee the existence of a binary linear code $\mathbf{C}$ of blocklength n, with at most c codewords in any ball of radius $e = pn$, and whose rate is $k = \lfloor (1 - H(p) - 1/c)n \rfloor$, for all large enough n. This clearly implies the lower bound on U_c^{const} claimed in the statement of the Theorem.

The code $\mathbf{C} = \mathcal{C}_k$ will be built iteratively in k steps by randomly picking the k basis vectors in turn. Initially the code $\mathcal{C}_0$ will just consist of the all-zeroes codeword $b_0 = 0^n$. The code $\mathcal{C}_i$, $1 \leq i \leq k$, will be successively built by picking a random (non-zero) basis vector b_i that is linearly independent of $b_1, \ldots, b_{i-1}$, and setting $\mathcal{C}_i = \text{span}(b_1, \ldots, b_i)$. Thus $\mathbf{C} = \mathcal{C}_k$ is an $[n, k]_2$ linear code. We will now analyze the list of c decoding radius of the codes $\mathcal{C}_i$, and the goal is to prove that the list of c decoding radius of $\mathbf{C}$ is at least e.

The key to analyzing the list of c decoding radius is the following potential function $S_{\mathcal{C}}$ defined for a code $\mathcal{C}$ of blocklength n:

$$S_{\mathcal{C}} = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} 2^{\frac{n}{c} \cdot |B(x,e) \cap \mathcal{C}|} . \tag{4}$$

For notational convenience, we denote $S_{\mathcal{C}_i}$ be S_i. Also denote by T_x^i the quantity $|B(x,e) \cap \mathcal{C}_i|$, so that $S_i = 2^{-n} \sum_x 2^{T_x^i}$.

Let $B = |B(0,e)|$; then $B \leq 2^{H(p)n}$ where $H(p)$ is the binary entropy function of p. Clearly

$$S_0 = 1 - B/2^n + B2^{n/c}/2^n \leq 1 + 2^{n\left(H(p) - 1 + 1/c\right)} . \tag{5}$$

Now once $\mathcal{C}_i$ has been picked with the potential function S_i taking on some value, say $\hat{S}_i$, the potential function S_{i+1} for $\mathcal{C}_{i+1} = \text{span}(\mathcal{C}_i \cup \{b_{i+1}\})$ is a random variable depending upon the choice of b_{i+1}. We consider the expectation $\mathbf{E}[S_{i+1}|S_i = \hat{S}_i]$ taken over the random choice of b_{i+1} chosen uniformly from outside $\text{span}(b_1, \ldots, b_i)$. It is not difficult to show that (see the full version of the paper for details), $\mathbf{E}[S_{i+1}|S_i = \hat{S}_i] = \hat{S}_i^2$ if the expectation is taken over all choices of b_{i+1} in $\{0,1\}^n$. We now use the simple fact that the expectation of a positive random variable taken over b_{i+1} chosen randomly from outside $\text{span}(b_1, \ldots, b_i)$ is at most $(1 - 2^{i-n})^{-1}$ times the expectation taken over b_{i+1} chosen uniformly at random from $\{0,1\}^n$. We thus get

$$\mathbf{E}[S_{i+1}|S_i = \hat{S}_i] \leq \frac{\hat{S}_i^2}{(1 - 2^{i-n})} \tag{6}$$

Applying (6) repeatedly for $i = 0, 1, \ldots, k-1$ and using the value of S_0 from Equation (5), we can show, after some calculations, that there exists an $[n, k]$ binary linear code $\mathbf{C}$ with

$$S_{\mathbf{C}} = S_k \leq 2(1 + 2 \cdot 2^{k+(H(p)-1+1/c)n}) \leq 6 \tag{7}$$

(the last inequality follows since $k = \lfloor(1-H(p)-1/c)n\rfloor$). By the definition of the potential S_k in Equation (4), this implies that $2^{n/c \cdot |B(x,e) \cap \mathbf{C}|} \leq 6 \cdot 2^n < 2^{n+3}$, or $|B(x,e) \cap \mathbf{C}| < (1 + \frac{3}{n})c$ for every $x \in \{0,1\}^n$. If $n > 3c$, this implies $|B(x,e) \cap \mathbf{C}| < c + 1$ for every x, implying that the list of c decoding radius of $\mathbf{C}$ is at least e, as desired. □ *(Theorem 4)*

Remark: It turns out that the statement of Theorem 4 can be adapted to guarantee the existence of certain (linear) codes which serve as excellent (for purposes of list decoding) inner codes in a concatenation scheme with an outer Reed-Solomon code. This in turn gives an *efficiently constructible* family of binary linear codes of rate ε^4 which can be efficiently list decoded from up to a $(\frac{1}{2} - O(\varepsilon))$ fraction of errors. This improves upon the results claimed in [7] (the best rate achieved by [7] for such families of codes was ε^6). Details of this construction will appear in the full version of this paper.

References

[1] V. M. Blinovskii. Bounds for codes in the case of list decoding of finite volume. *Prob. Information Transmission*, **22** (1), pp. 11-25 (in Russian), 1986; pp. 7-19 (in English), 1986.

[2] I. Dumer, D. Micciancio and M. Sudan. Hardness of approximating the minimum distance. *Proc. of STOC 1999.*

[3] P. Elias. List decoding for noisy channels. *Wescon Convention Record*, Part 2, Institute of Radio Engineers (now IEEE), pp. 94-104, 1957.

[4] P. Elias. Error-correcting codes for List decoding. *IEEE Trans. Info. Theory*, **37** (1), pp. 5-12, 1991.

[5] O. Goldreich, R. Rubinfeld and M. Sudan. Learning polynomials with queries: the highly noisy case. *Proc. of FOCS 95.*

[6] V. Guruswami and M. Sudan. Improved decoding of Reed-Solomon and Algebraic-geometric codes. *IEEE Trans. on Information Theory*, 45 (1999), pp. 1757-1767. Preliminary version appeared in *Proc. of FOCS'98.*

[7] V. Guruswami and M. Sudan. List decoding algorithms for certain concatenated codes. *Proc. of STOC 2000.*

[8] G. H. Hardy, J. E. Littlewood, G. Pólya. *Inequalities*, 2nd Edition, Cambridge University Press, 1952.

[9] J. Justesen and T. Hoholdt. Bounds on list decoding of MDS codes. Preprint, 1999.

[10] C. E. Shannon, R. G. Gallager and E. R. Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels. *Information and Control*, **10**, pp. 65-103 (Part I), pp. 522-552 (Part II), 1967.

[11] M. A. Shokrollahi and H. Wasserman. List decoding of algebraic-geometric codes. *IEEE Trans. on Information Theory*, Vol. 45, No. 2, March 1999, pp. 432-437.

[12] M. Sudan. Decoding of Reed-Solomon codes beyond the error-correction bound. *Journal of Complexity*, 13(1):180-193, March 1997.

[13] A. Ta-Shma and D. Zuckerman. Personal Communication, April 1999.

[14] J. H. van Lint. *Introduction to Coding Theory*, Graduate Texts in Mathematics **86**, (Third Edition) Springer-Verlag, Berlin, 1999.

[15] J. M. Wozencraft. List Decoding. *Quarterly Progress Report*, Research Laboratory of Electronics, MIT, Vol. 48 (1958), pp. 90-95.

[16] V. V. Zyablov and M. S. Pinsker. List cascade decoding. In *Prob. Information Transmission*, **17** (4), pp. 29-34 (in Russian), 1981; pp. 236-240 (in English), 1982.

The Displacement Method in Coding Theory

Vadim Olshevsky
Department of Mathematics
and Computer Science
Georgia State University
Atlanta, GA
volshevsky@gsu.edu

Amin Shokrollahi
Digital Fountain, Inc.
600 Alabama Street
San Francisco, CA 94110
amin@digitalfountain.com

September 28, 2000

Abstract

For many classes of algebraic codes, encoding and decoding reduces to computing the product of a structured matrix with a vector, and finding a nontrivial element in the kernel of a structured matrix. We will use the method of displacement to attack these problems. Specifically, we concentrate on the problem of efficiently finding nonzero elements in the kernel of a structured matrix. The concept of structure is formalized using the displacement operator. The structure of such an $n \times n$- matrix allows it to be "compressed" to αn parameters for some α which is usually much smaller than n. The displacement operator allows to perform certain matrix operations like the PLU-decomposition on the compressed version, thereby reducing the running time from cubic to quadratic. As examples of our methods we will derive efficient sequential and parallel algorithms for list decoding of Reed-Solomon codes.

1 Introduction

Linear error-correcting codes are an indispensable tool in various communication media. They are used in scenarios where information is transmitted over a channel and can potentially be subjected to corruption. The idea of error-correcting codes is to add redundancy to the original message so as to be able to correct errors.

In algebraic coding theory one typically assumes that the elements of the transmission alphabet form a finite field $\mathbb{F}_q$. The most common case is when q is a power of 2, in which case the elements of the alphabet can be represented as binary strings. A *linear code* of block-length n and dimension k is then a k-dimensional subspace of $\mathbb{F}_q^n$. Here and in the following we assume that $\mathbb{F}_q^n$ is equipped with the standard basis, so that elements of this space can be represented as n-tuples over $\mathbb{F}_q$. The *Hamming distance* between two elements in $\mathbb{F}_q^n$ is defined as the number of coordinate places at which they differ. The minimum distance of a code is the minimum Hamming distance between two distinct codewords. A code of block-length n, dimension k, and minimum distance d over $\mathbb{F}_q$ is denoted as an $[n, k, d]_q$-code. If the number of errors in a transmission is at most $(d-1)/2$, then it is possible to correct the errors by searching for the (unique) codeword of minimum Hamming distance to the received word. This algorithm is not very efficient since it requires in the worst case visiting all the codewords of which there are q^k many. The aim is to design more efficient algorithms for special classes of codes.

As an example, consider the Reed-Solomon codes, one of the most important classes of codes both from a theoretical and a practical point of view [16]. Fix n different elements $x_1, \ldots, x_n$ of the field $\mathbb{F}_q$, and suppose that $k \leq n$ are positive integers. The image of the homomorphism γ mapping a polynomial f of degree less than k to the vector $(f(x_1), \ldots, f(x_n))$ is a linear subspace of $\mathbb{F}_q^n$, hence a linear code. Using the fact that the number of zeros over $\mathbb{F}_q$ of a nonzero polynomial is at most equal to its degree on can easily see that this code has dimension k and minimum distance $d = n-k+1$. Encoding of this code is easy: it is essentially the multiplication of a Vandermonde matrix given by the x_i with a vector given by the coefficients of the polynomial f. This task can either be solved in time $O(n^2)$ using standard algorithms, or in time $O(n \log^2(n) \log\log(n))$ using fast algorithms [3, Chap. 2]. The decoding is slightly more complicated. Here we give a conceptually simple algorithm due to Welch and Berlekamp [15, 2]: suppose that we have received a word $y = (y_1, \ldots, y_n)$ and we have a promise that there is a polynomial f such that the Hamming distance of $(f(x_1), \ldots, f(x_n))$ and y is at most $(d-1)/2 = (n-k)/2$. The decoding problem is that of efficiently computing f. In a first step, compute polynomials g and h, not both zero, of degrees $< (n+k)/2$ and $\leq (n-k)/2$, respectively, such that $g(x_i)+y_i h(x_i) = 0$ for $i = 1, \ldots, n$. Then g is divisible by h and the desired polynomial f equals $f = -g/h$. This statement is easily proved: the polynomial $H = g+fh$ has degree less than $(n+k)/2$ and vanishes at all x_i such that $f(x_i) = y_i$. By assumption, there are at least $n-(n-k)/2 = n+k/2$ such i's. Hence, H has more zeros than its degree. So $H = 0$ and $f = -g/h$. But how to we compute g and h? This is done by considering the coefficients of g and h as unknowns and regarding each of the conditions $g(x_i)+y_i h(x_i) = 0$ as a linear condition on these unknowns. There are n conditions, and there are $n+1$ unknowns (at least when $n \equiv k \bmod 2$). Hence, the vector of coefficients of (g, h) is a nonzero element in the kernel of a matrix and can be computed using Gaussian elimination in time $O(n^3)$. But note that the matrix we are dealing with is structured. The method of displacement can now be used to compute a PLU-decomposition of the matrix and compute a kernel element in time $O(n^2)$.

The above example is by no means isolated. There are many other examples of types of structure which are directly related to the decoding problem for certain algebraic codes. For instance, Hankel matrices are encountered in the Berlekamp-Massey algorithm [1], and Cauchy matrices related to decoding classical Goppa-codes [8]. Existing algorithms make use of the structure of these matrices, but the methods are usually confined to one particular type of structure. On the contrary, we will in this paper introduce the displacement method which can be thought of as a generator of efficient algorithms if certain mild criteria on the underlying structure are fulfilled. Though this method has been successfully used in other contexts such as image processing, system theory, or interpolation [7, 9, 10], its use in coding theory is novel and quite powerful. Its power stems from the fact that it enables one to perform matrix operations on "compressed" versions of a structured matrix, rather than on the matrix itself. Note that Gaussian elimination on an $n \times n$-matrix requires $O(n^3)$ time since the entries of the matrix have to be updated n times. Hence, if the compressed form of the matrix has significantly less elements, and each if round of the algorithm updates only the entries of the compressed version, then we obtain an algorithm that uses significantly fewer operations.

To exemplify applications of our approach we will design efficient parallel and sequential algorithms for list decoding algorithms for Reed-Solomon codes of which the above decoding algorithm is a special case.

The specific decoding problems highlighted in this paper only serve as examples of the power of the displacement method in coding theory, and comprise by no means a complete list. The last section of this paper describes further coding theoretic problems that can be successfully attacked using our method.

2 The Displacement Structure

2.1 Sequential Algorithms

Let m and n be positive integers, K be a field, and $D \in K^{m \times m}$, $A \in K^{n \times n}$. We define the *displacement operator* $\nabla = \nabla_{D,A}\colon K^{m \times n} \to K^{m \times n}$ by $\nabla(V) := DV - VA$. The ∇-*displacement rank* of V is defined as the rank of the matrix $\nabla(V)$. If this rank is α, then $\nabla(V)$ can be written as GB with $G \in K^{m \times \alpha}$ and $B \in K^{\alpha \times n}$. The pair (G, B) is then called a ∇-*generator* for V. If α is small and D and A are sufficiently simple, then the ∇-operator allows to "compress" the matrix V to matrices with a total of $\alpha(m+n)$ entries. Furthermore, one can efficiently compute with the compressed form as the following lemma suggests.

Lemma 1. *Let the matrices in $DV - VA = GB$ be partitioned as*

$$G = [g_{11} \mid G_{21}],\ B = [b_{11} \mid B_{12}],\ D = \begin{bmatrix} d_1 & 0, \\ \star & D_2 \end{bmatrix} A = \begin{bmatrix} a_1 & \star \\ 0 & A_2 \end{bmatrix},\ V = \begin{bmatrix} v_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

Suppose that v_{11} is nonzero. Then the Schur complement $V_2 := V_{22} - v_{11}^{-1} V_{21} V_{12}$ of V satisfies the equation

$$D_2 V_2 - V_2 A_2 = G_2 B_2,$$

where $G_2 = G_{21} - v_{11}^{-1} g_{11} V_{12}$ and $B_2 = B_{12} - v_{11}^{-1} b_{11} V_{12}$.

A proof can be found in [11, Lemma 3.1], see also [5]. The lemma shows how to obtain an LU-factorization for V (if it exists), by operating only on the matrices G and B. For this, one only needs to know the following well-known facts: the first column of L below the upper left entry is given by the first column of the Schur-complement V_2, and the first row of U is given by the first row of V_2.

The lemma suggests an algorithm for computing an LU-decomposition of V which may not work if the upper left entry of V of any of its successive Schur complements are zero. In this situation partial pivoting can be used. This leads to a modification of the above lemma in which we have to assume that the matrix D is diagonal. We briefly sketch the modification; details can be found in [10, Sect. 3.5]. If $v_{11} = 0$ and the first column of V contains a nonzero element, say at position $(k, 1)$, then we can consider PV instead of P, where P is the matrix corresponding to interchanging rows 1 and k. The displacement equation for PV is then given by $(PDP^\top)(PV) - PVA = PGB$. Since P can be an arbitrary transposition, for $PDP^\top$ to be lower triangular we have to assume that D is diagonal. This explains the assumptions in the algorithm below (see [10, Sect. 3.5]).

Algorithm 2. *On input a diagonal matrix $D \in K^{m \times m}$, an upper triangular matrix $A \in K^{n \times n}$, and a $\nabla_{D,A}$-generator (G, B) for $V \in K^{m \times n}$ in $DV - VA = GB$, the algorithm outputs a permutation matrix P, a lower triangular matrix $L \in K^{m \times m}$, and an upper triangular matrix $U \in K^{m \times n}$, such that $V = PLU$.*

(1) *Recover from the generator the first column of V.*

(2) *Determine the position, say $(k,1)$, of a nonzero entry of V. If it does not exist, then set the first column of L equal to $[1,0]^\top$ and the first row of U equal to the first row of V, and go to Step (4). Otherwise interchange the first and the k-th diagonal entries of A and the first and the k-th rows of G and call P_1 the permutation matrix corresponding to this transposition.*

(3) *Recover from the generator the first row of $P_1V =: \left(\begin{smallmatrix} v_{11} & V_{12} \\ V_{21} & V_{22} \end{smallmatrix}\right)$. Store $[1, V_{12}/v_{11}]^\top$ as the first column of L and $[v_{11}, V_{12}]$ as the first row of U.*

(4) *If $v_{11} \neq 0$, compute by Lemma 1 a generator of the Schur complement V_2 of P_1V. If $v_{11} = 0$, then set $V_2 := V_{22}$, $G_2 := G_{21}$, and $B_2 := B_{12}$.*

(5) *Proceed recursively with V_2 which is now represented by its generator (G_2, B_2) to finally obtain the factorization $V = PLU$, where $P = P_1 \cdots P_\mu$ with P_k being the permutation used at the k-th step of the recursion and $\mu = \min\{m,n\}$.*

The correctness of the above algorithm and its running time depend on steps (1) and (3). Note that it may not be possible to recover the first row and column of V from the matrices D, A, G, B. In fact, recovery from these data alone is only possible if $\nabla_{D,A}$ is an isomorphism. For simplicity we assume in the following that this is the case. In the general case one has to augment the (D, A, G, B) by more data corresponding to the kernel of ∇, see [10, Sect. 5].

Lemma 3. *Suppose that steps (1) and (3) of Algorithm 2 run in time $O(m)$ and $O(n)$, respectively. Then the total running time of that algorithm is $O(\alpha mn)$, where α is the displacement rank of V with respect to $\nabla_{D,A}$.*

Proof. The proof is obvious once one realizes that Step (4) runs in time $O(\alpha(m+n))$, and that the algorithm is performed recursively at most $\min\{m,n\}$ times. ☐

In this paper we are mainly concerned with finding a nonzero element in the kernel of V. Once a PLU-decomposition for V is known, such an element can be found in time $O(\min\{m,n\}^2)$ by the straightforward backward substitution algorithm for solving a homogeneous upper triangular system of equations.

Corollary 4. *Suppose that the matrix $V \in K^{m\times n}$ has displacement rank α with respect to the isomorphism $\nabla_{D,A}$ and that V has rank less than n. Then one can compute a nonzero element in the kernel of V with $O(\alpha mn)$ operations.*

2.2 Parallel Algorithms

The algorithm given in the last subsection can be customized to run in parallel if certain additional conditions are satisfied. We will sketch the approach here.

To begin with, assume that the matrix A of Algorithm 2 is equal to D. This will enable performing steps (1) and (3) of that algorithm in parallel, as we will see below. In this case, however, the operator $\nabla_{D,D}$ is not an isomorphism anymore, so that additional work is necessary to recover the matrix V from the data D, G, B.

The additional data consists of the diagonal entries $v_{11}, v_{22}, \ldots, v_{nn}$ of the matrix V. During the course of the algorithm these values are updated to the diagonal entries of the Schur-complements.

The first column of V is obtained in the following way: first, compute the first column $(\gamma_1, \ldots, \gamma_m)^\top$ of GB. Using m processors, this takes $O(\alpha)$ time, where α is the displacement rank of V with respect to $\nabla_{D,D}$. This vector equals the first column of $DV - VD$, i.e., it equals $(0, (d_2 - d_1)v_{21}, \ldots, (d_m - d_1)v_{m1})^\top$, where we have denoted the diagonal entries of D by $d_1, \ldots, d_m$. From this, one can compute the first column of V with m processors in constant time. If the entry v_{11} is nonzero, then one can in the same way as above compute the first row of V and proceed with the algorithm. Additional care has to be taken, however, when $v_{11} = 0$. The somewhat lengthy, but straightforward details will be presented in the final version of the paper.

The algorithm then proceeds exactly in the same way as Algorithm 2 to obtain a PLU-decomposition. Since solving a homogeneous upper triangular system of linear equations can be easily customized to run in parallel linear time, this gives an algorithm for computing a nontrivial element in the kernel of V in time $O(\alpha n)$ on $O(n)$ processors.

3 List Decoding of Reed-Solomon Codes

Sudan [14] has designed an algorithm for *bounded distance list decoding* of Reed-Solomon codes which we will briefly recall for the benefit of the reader. Given a vector $y = (y_1, \ldots, y_n)$ in $\mathbb{F}_q^n$ and a promise that the Hamming distance of this vector and the Reed-Solomon code C is most e (for some appropriate e), the algorithm returns the list of all codewords of Hamming distance $\leq e$ from y. For the convenience of the reader we will recall the algorithm. Let $x_1, \ldots, x_n$ be different elements in $\mathbb{F}_q$ and let C be the k-dimensional Reed-Solomon code of block-length n over $\mathbb{F}_q$ whose elements are of the form $(f(x_1), \ldots, f(x_n))$ where f is a polynomial over $\mathbb{F}_q$ of degree less than k. Fix a positive integer ℓ (the list length) and set

$$e := \left\lfloor \frac{\ell}{\ell+1} n - \frac{k\ell}{2} \right\rfloor, \quad b := \left\lceil \frac{n}{\ell+1} + \frac{k\ell}{2} + 1 \right\rceil.$$

Given $y = (y_1, \ldots, y_n)$, the algorithm computes in a first step polynomials $g_0, \ldots, g_\ell$, not all 0 and $\deg(g_i) < b - ik$, such that

$$\forall\, i = 1, \ldots, n: \qquad g_0(x_i) + y_i g_1(x_i) + \cdots + y_i^\ell g_\ell(x_i) = 0. \tag{1}$$

Then all codewords of distance $\leq e$ from y are given as evaluations of the polynomials f such that

$$g_0(x) + f(x) g_1(x) + \cdots + f(x)^\ell g_\ell(x) = 0.$$

There exist efficient algorithms for computing the polynomials f satisfying the above equality [4]. We will thus concentrate on the problem of computing $g_0, \ldots, g_\ell$. As in the case of conventional decoding of Reed-Solomon codes outlined in the introduction, this is achieved by solving a homogeneous system of linear equations. The corresponding matrices have a highly repetitive pattern, and the following notation will help us to concisely describe them. Let $\varphi, \varphi_1, \ldots, \varphi_t \colon M \to K$ be functions from a set M into a field K, and let $m_1, \ldots, m_n$ be elements in M. We define

$$[\varphi_1(m), \ldots, \varphi_t(m)]_{(m_1,\ldots,m_n)} := \begin{pmatrix} \varphi_1(m_1) & \varphi_2(m_1) & \cdots & \varphi_t(m_1) \\ \varphi_1(m_2) & \varphi_2(m_2) & \cdots & \varphi_t(m_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(m_n) & \varphi_2(m_n) & \cdots & \varphi_t(m_n) \end{pmatrix}.$$

Whenever $m_1, \ldots, m_n$ are clear from the context, we will replace the subscript $(m_1, \ldots, m_n)$ by $1, n$. Further, we denote the diagonal matrix with diagonal entries $\varphi(m_1), \ldots, \varphi(m_n)$ by $\operatorname{diag}[\varphi]_{1,n}$.

The polynomial $F(x, y) := g_0(x) + y g_1(x) + \cdots + y^\ell g_\ell(x)$ corresponds to a nonzero element in the kernel of a matrix V with a repetitive structure which we will describe below. Let $d_0, \ldots, d_\ell \geq 0$ be defined by $d_i := b - ik - 1$. For each $i = 0, \ldots, b$ let V_i be the matrix

$$V_i := \left[y^i x^{d_i}, \cdots, y^i x, y^i \right]_{1,n}$$

and define the matrix $V \in K^{n \times (n+1)}$ with the block decomposition

$$V = (V_\ell \mid V_{\ell-1} \mid \cdots \mid V_1 \mid V_0) \,. \tag{2}$$

(Here and for the rest of this section the subscript "$1, n$" will always refer to the points $(x_1, y_1), \ldots, (x_n, y_n)$.) Let $\mathfrak{x}$ be a nonzero vector such that $V\mathfrak{x} = 0$. Then, interpreting the entries of $\mathfrak{x}$ as coefficients of the g_j (starting from g_ℓ down to g_0 and reading the coefficients from high powers to low powers of x), this gives a polynomial F as desired. Using the displacement structure approach, we can easily develop an algorithm with running time $O(\ell n^2)$ for computing F. Assuming that ℓ is a constant (a reasonable assumption in applications), this gives an algorithm with a running time that is quadratic in n.

For computing F we will first prove that V has displacement rank at most $\ell + 1$. Let $D := \operatorname{diag}[x]_{1,n}$ (i.e., D is the diagonal matrix with diagonal entries $x_1, \ldots, x_n$). Further, let $A \in K^{(n+1) \times (n+1)}$ be the *upper shift matrix* of format $n + 1$, i.e.,

$$A := \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

Let

$$G = \left[y^b x^{d_\ell}, y^{\ell-1}(x^{d_{\ell-1}} - y), \cdots, y(x^{d_1} - y), x^{d_0} - y \right]_{1,n} \in K^{n \times (\ell+1)} \tag{3}$$

and

$$B = (B_\ell \mid B_{\ell-1} \mid \cdots \mid B_1 \mid B_0) \in K^{(\ell+1) \times (n+1)}, \tag{4}$$

where each B_i is an $(\ell+1) \times (d_i+1)$-matrix whose only nonzero position is at $(\ell+1-i, 1)$. Then, one easily proves that

$$DV - VA = GB. \tag{5}$$

Hence, V has displacement rank at most $\ell + 1$ with respect to $\nabla_{D,A}$. We can now invoke the general algorithm 2 to compute a PLU-decomposition of V using only the matrices D, G, B. (In fact, since we are only interested in nonzero elements in the kernel of V, it suffices to compute the matrix U only.) To apply Algorithm 2 we need to recover the first row and the first column of V from the first row and the first column of GB. The next algorithm shows how to do this in the case where all the x_i are nonzero (which is equivalent to $\nabla_{D,A}$ being an isomorphism).

Algorithm 5. *On input a diagonal matrix D with diagonal entries $x_1, \ldots, x_m$, all of them nonzero, a matrix $G \in K^{m \times \alpha}$ as in (3), and a matrix $B \in K^{\alpha \times (m+1)}$ as in (4), this algorithm computes the first row and the first column of the matrix V given by $DV - VA = GB$, where A is the upper shift matrix of format $m + 1$.*

(1) *Compute the first row* $(r_1, \ldots, r_{m+1})$ *and the first column* $(\gamma_1, \ldots, \gamma_n)^\top$ *of* GB.

(2) *For* $i = 1, \ldots, m$ *set* $v_i := \gamma_i / x_i$.

(3) *Set* $c_1 := v_1$. *For* $i = 2, \ldots, m+1$ *set* $c_i := (r_i + c_{i-1})/x_i$.

(4) *Output row* $(c_1, \ldots, c_{m+1})$ *and column* $(v_1, \ldots, v_m)^\top$.

Proposition 6. *The above algorithm correctly computes its output with* $O(dm)$ *operations in the field* K.

Proof. Correctness of the algorithm follows immediately from the following observation: let $(c_1, \ldots, c_{m+1})$ and $(c_1, v_2, \ldots, v_m)^\top$ be the first row and the first column of V, respectively. Then

$$DV - VA = \begin{pmatrix} x_1c_1 & x_1c_2 - c_1 & \cdots & x_1c_{m+1} - c_m \\ x_2v_2 & \star & \cdots & \star \\ \vdots & \vdots & \cdots & \vdots \\ x_mv_m & \star & \cdots & \star \end{pmatrix},$$

where the $\star$'s denote elements in K. As for the running time, observe first that computing the first row and first column of GB requires at most $2dm$ operations over K. Further, Step (2) and (3) require at most m and $2m$ operations, respectively. □

Using the above result we obtain the following general algorithm for list decoding of Reed-Solomon codes for the case where none of the x_i is zero.

Algorithm 7. *On input* $(x_1, y_1), \ldots, (x_n, y_n) \in K^2$ *and integers* $d_0 \geq d_1, \ldots, d_\ell$ *such that* $\sum_{i=0}^{\ell}(d_i + 1) = n + 1$ *and such that none of the* x_i *is zero, this algorithm computes a nonzero solution of the homogeneous system of linear equations given by* $V\mathfrak{x} = 0$, *where* V *is given in (2).*

(1) *Set* $G_1 := G$, $B_1 := B$, $U := 0$, *where* G *and* B *are given in (3) and (4) respectively.*

(2) *For* $i = 1, \ldots, n$ *do*

   (a) *Let* D_i *be the diagonal matrix with diagonal entries* $x_i, \ldots, x_n$. *Use Algorithm 5 with input* D_i, G_i, *and* B_i *to compute column* $\mathfrak{c} = (c_1, \ldots, c_{n-i+1})^\top$.

   (b) *If* $\mathfrak{c}$ *is nonzero, then find location* k *such that* $c_k \neq 0$. *Interchange rows* k *and* 1 *of* $\mathfrak{c}$ *and of* G_i. *If* $\mathfrak{c}$ *is zero, then perform the next step, increment* i *by* 1 *and go to Step* (a).

   (c) *Use Algorithm 5 with input* D_i, G_i *and* B_i *to compute row* $\mathfrak{r} = (r_1, \ldots, r_{n+2-i})$. *Write* $\mathfrak{r}$ *in* i*-th row of matrix* U *starting at position* i.

   (d) *For* $j = 2, \ldots, n-i+1$ *replace row* j *of* G_i *by* $-c_j/c_1$ *times the first row plus the* j*-th row. Set* G_{i+1} *as the matrix formed from* G_i *by deleting the first row.*

   (e) *For* $j = 2, \ldots, n-i+2$ *replace the* j*-th column of* B_i *by* $-r_j/c_1$ *times the first column plus the* j*-th column. Set* B_{i+1} *as the matrix formed from* B_i *by deleting the first column.*

(3) *Find nonzero solution of* $U\mathfrak{x} = 0$ *and output* $\mathfrak{x}$.

Theorem 8. *The above algorithm correctly computes its output with $O(\ell n^2)$ operations over the field K.*

PROOF. Follows at once from Corollary 4 and Proposition 6. □

The performance of the algorithm can be further enhanced by the following observation: since we are only interested in a nonzero element in the kernel of V, we can stop the algorithm, as soon as we find a zero element in the diagonal of the matrix U. A zero element on the main diagonal of U is obtained after at most $\mathrm{rk}(V)$ steps. Hence, the running time of the algorithm with this "early abort" strategy is $O(n\ell \mathrm{rk}(V))$.

We further remark that the algorithm can be modified to deal with the case that one x_i is zero as well. One way to carry this out is described below. We may w.l.o.g. assume that $x_1 = 0$. In this case, deletion of the first row of the matrix V yields another matrix $\tilde{V}$ of the form given in (2) in which none of the x_i is zero. Algorithm 7 computes for $\tilde{V}$ an upper triangular matrix $\tilde{U}$ such that $\tilde{V} = \tilde{P}\tilde{L}\tilde{U}$ for some permutation matrix $\tilde{P}$ and some nonsingular lower triangular matrix $\tilde{L}$. Let U be the matrix whose first row is that of V, and whose remaining rows are those of $\tilde{U}$. It is then obvious that there is a permutation matrix P and a nonsingular lower triangular matrix L such that $V = PLU$. Hence, a vector $\mathfrak{x}$ satisfies $V\mathfrak{x} = 0$ iff it satisfies $U\mathfrak{x} = 0$. Because of the structure of U, one can find such a nonzero vector $\mathfrak{x}$ in time $O(n^2)$.

The displacement operator described in (5) will not allow the design of parallel algorithms, since the matrix A is not diagonal. To remedy this situation, we use the following technique:

Proposition 9. *Let the matrices $V \in K^{m\times n}$ and $W \in K^{n\times \ell}$ have displacement ranks r_1 andr_2 with respect to $\nabla_{F,A}$ and $\nabla_{A,R}$, respectively. Then VW has displacement rank at most $r_1 + r_2$ with respect to $\nabla_{F,R}$. Moreover, a generator for VW can be obtained from the generators of V and W with $O((r_1 + r_2)n^2)$ sequential time, and in time $O((r_1+_2)n)$ on $O(n)$ processors.*

The proof of this result follows trivially from the definitions. The assertion on the running time follows from the sequential and parallel running times of the trivial matrix multiplication algorithms.

To obtain an efficient parallel algorithm for computing an element in the kernel of the matrix V defined in (2), choose $n + 1$ distinct elements from $\mathbb{F}_q$ (or an extension thereof) and denote by W the Vandermonde matrix corresponding to these elements and by Δ the diagonal matrix having these elements as its diagonal entries. Further, let C denote the upper shift matrix of format $n + 1$, and let D denote the diagonal matrix having entries $x_1, \ldots, x_n$. As usual, we assume that D and Δ are invertible. Since V has displacement rank $\leq \ell + 1$ with respect to $\nabla_{D,C}$ (see (5)), and W has displacement rank one with respect to $\nabla_{\Delta,C^\top}$, Proposition 9 proves that $WV^\top$ has displacement rank $\leq \ell + 2$ with respect to $\nabla_{\Delta,D}$, and that generators of this operator can be calculated in time $O(\ell n)$ on $O(n)$ processors. Using results of Section 2.2, we see that we can compute a PLU-decomposition of $WV^\top$ in time $O(\ell n)$ on $O(n)$ processors. It is now easy to see that from this we can compute a nontrivial element in the kernel of V in time $O(n)$ on $O(n)$ processors. The final algorithm is a parallel algorithm that computes a nontrivial element in the kernel of the matrix V in time $O(\ell n)$ on $O(n)$ processors.

4 The Improved Algorithm

In [6] the authors describe an extension of algorithm presented in [14] in which they use a variant of Hermite interpolation rather than a Lagrange interpolation. In the case of Reed-Solomon codes, the input to the algorithm is a set of n points (x_i, y_i) in the affine plane over $\mathbb{F}_q$, and parameters r, k, and ℓ. Let β be the smallest integer such that $(\ell+1)\beta > \binom{\ell+1}{2}k + \binom{r+1}{2}n$. The output of the algorithm is a nonzero bivariate polynomial $G = \sum_{i=0}^{\ell} G_i(x)y^i$ such that $\deg G_i < \beta - ik$, $G(x_i, y_i) = 0$ for all $i \le n$, and $G(x+x_i, y+y_i)$ does not contain any monomial of degree less than r. Such a polynomial G corresponds to a nonzero element in the kernel of a certain matrix V which we will describe below. Let $t \le n$ be a fixed positive integer. For $0 \le i < r$ and $0 \le j \le \ell$ let $V_{ij}^t \in \mathbb{F}_q^{(r-i)\times(\beta-jk)}$ be the matrix having (μ,ν)-entry equal to $\binom{\nu}{\mu}x_t^{\nu-\mu}$, where $0 \le \mu < r-i$ and $0 \le \nu \le \beta - jk$. Now define the matrix V_t as a block matrix with r block rows and $\ell+1$ block columns, with block entry (i,j) equal to $\binom{j}{i}y_t^{j-i}V_{ij}^t$. The matrix V then equals

$$V = \begin{pmatrix} V_1 \\ V_2 \\ \vdots \\ V_n \end{pmatrix}. \tag{6}$$

V has $m := \binom{r+1}{2}n$ rows and $s := (\ell+1)\beta - \binom{\ell+1}{2}k$ columns (and $s > m$). Let C be the lower shift matrix of format s, i.e., C is the transpose of the upper shift matrix of format s introduced in Section 3. Further, let J_i^t denote the $i \times i$-Jordan block having x_t in its main diagonal and 1's on its lower sub-diagonal. Let J^t be the block diagonal matrix with block diagonal entries $J_r^t, \ldots, J_1^t$. Let J be the block diagonal matrix with block diagonal entries $J^1, \ldots, J^n$. Then, a quick calculation shows that $JV - VC$ has rank at most $\ell + 1$. It is now tempting to apply the general results of Section 2 to this situation. However, J is not a diagonal matrix. Therefore, we will use Proposition 9.

Let $\mathbb{F}$ be a suitable extension of $\mathbb{F}_q$ having at least s elements, and denote by W the $(s \times s)$-Vandermonde matrix whose rows consist of powers of these elements. Let Δ denote the diagonal matrix having these elements as its diagonal entries. To avoid tedious arguments, we assume that none of the diagonal entries of Δ are zero. Then W has displacement rank one with respect to $\nabla_{\Delta,C^\top}$. Thus, the previous proposition shows that $WV^\top$ has displacement rank $\le \ell + 2$ with respect to $\nabla_{\Delta,J^\top}$. We can now apply Algorithm 7 to obtain a PLU-decomposition of $WV^\top$, where $P, L \in \mathbb{F}^{s\times s}$ and $U \in \mathbb{F}^{s\times m}$. By Corollary 4 and Proposition 6 this takes $O(\ell sm)$ operations over the field $\mathbb{F}$. Further, we obtain $V = U^\top P^\top L^\top (W^{-1})^\top$. To find a nontrivial element v in the kernel of V, we first compute a nontrivial element u in the kernel of $U^\top$; this can be achieved with $O(m^2)$ operations over the field $\mathbb{F}$. Next we solve the system of linear equations $L^\top w = Pu$. Since $L^\top$ is upper triangular and of full rank m, this takes $O(m^2)$ operations. The desired element v is then obtained as $v = W^\top w$, and its computation takes $O(s^2)$ operations. In total, this gives a sequential algorithm with running time $O(s^2\ell)$ over the field $\mathbb{F}$. Each operation in $\mathbb{F}$ uses $O(\log_q^2(s))$ operations over the base field $\mathbb{F}_q$. Hence, we obtain an algorithm with running time $O(s^2 \log_q^2(s)\ell)$. In the algorithm of Guruswami and Sudan [6] s equals $O(r^2n)$. Furthermore n and q have the same order of magnitude. As a result, we obtain an algorithm with running time $O(n^2r^4 \log_q(r)\ell)$. In many practical situations r and ℓ are constant; hence this gives an algorithm with running time $O(n^2)$.

5 Reducing the memory requirements

Above we designed algorithms for computing a PLU-decomposition for a structured matrix. From this we could compute a nonzero element in the kernel of the original matrix by solving an upper triangular homogeneous system. However, considering that we are only interested in a vector of length n, say, and that we have to keep track of a matrix U of considerably larger size $O(n^2)$, this method seems to waste a lot of memory. By applying the displacement method to a modified matrix, we will show in this section how to solve our original problem efficiently, both in terms of time and space.

The method we will present will find the smallest "column dependency" of a matrix V with appropriate displacement structure. By this we mean the computation of the minimal t such that the first $t+1$ columns of V are linearly dependent, and the computation of the coefficients of this linear dependency. In other words, we compute the minimal t and a nonzero vector in the kernel of V which has zeros in coordinate positions $t+2, \ldots, n$, and a one in position $t+1$. The algorithm is based on the following observation: let I_s denote the $s \times s$-identity matrix, and suppose that $V \in \mathbb{F}_q^{m \times n}$. Consider the matrix

$$\begin{pmatrix} V \\ I_n \end{pmatrix} =: \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \\ I_t & 0 \\ 0 & I_{n-t} \end{pmatrix}.$$

Assume that V_{11} is nonsingular. Since the first t columns of V are linearly independent by assumption, this will always be the case of we premultiply V with an appropriate permutation matrix. The Schur-complement of the new matrix above with respect to V_{11} is

$$\begin{pmatrix} V_{22} - V_{21}V_{11}^{-1}V_{12} \\ -V_{11}^{-1}V_{12} \\ I_{n-t} \end{pmatrix}.$$

Note that the first entry in the above matrix is the Schur-complement of V with respect to V_{11} whose first column is zero since the first $t+1$ columns of V are assumed to be dependent. We have

$$V\begin{pmatrix} -V_{11}^{-1}V_{12} \\ I_{n-t} \end{pmatrix} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}\begin{pmatrix} -V_{11}^{-1}V_{12} \\ I_{n-t} \end{pmatrix} = \begin{pmatrix} 0 \\ V_{22} - V_{21}V_{11}^{-1}V_{12} \end{pmatrix}$$

which shows that the first column of the matrix $\binom{-V_{11}^{-1}V_{12}}{I_{n-t}}$ is a linear dependency of the first $t+1$ columns of V. Hence, if $\binom{V}{I_n}$ has an appropriate displacement structure, we can run Algorithm 2 on this matrix, discarding the information on U, and only checking whether at the ith iteration the first $m-i-1$ entries of the first column of the current matrix are zero. In that case, we can easily recover a kernel element from the remaining entries of that column.

To show that $\binom{V}{I_n}$ has a displacement structure, suppose that $DV - VA = GB$, with a diagonal matrix D and an upper triangular matrix A. Then we have

$$\begin{pmatrix} D & 0 \\ 0 & A \end{pmatrix}\begin{pmatrix} V \\ I_n \end{pmatrix} - \begin{pmatrix} V \\ I_n \end{pmatrix} A = \begin{pmatrix} G \\ 0 \end{pmatrix} (B \mid 0).$$

The following observation shows that the first matrix above does not cause pivoting problems even though it is not diagonal: first, note that the algorithm runs for at most

m steps. Next, note that one never interchanges one of the first m rows of the matrix with one of the remaining n rows: if that becomes necessary, then the first column of the corresponding Schur-complement of V is zero and the algorithm is finished. Hence, pivoting takes place only among the first m rows, and this does not cause problems since this only leads to a simultaneous change of columns and rows of the diagonal part of the matrix; hence, this part will always remain diagonal.

A similar procedure can be developed when looking for a nonzero element x such that $xV = 0$.

6 Conclusions

In this paper we have introduced a general method originating from numerical analysis for efficient sequential and parallel list decoding of Reed-Solomon codes. Our algorithm computes a PLU-decomposition of a given dense structured $(n \times m)$-matrix in time close to $O(n^2)$, where closeness depends on the so-called displacement rank of the matrix. We also sketched how to modify our algorithms to use a much smaller amount of memory. Our methods can also be applied to derive efficient algorithms for list decoding of Algebraic-Geometric codes [13, 12]. There are many more applications of the displacement approach to coding theoretic problems, some of which will appear elsewhere.

References

[1] E. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[2] E. Berlekamp, "Bounded distance + 1 soft decision Reed-Solomon decoding," *IEEE Trans. Inform. Theory* **42**, pp. 704–720, 1996.

[3] P. Bürgisser, M. Clausen, and A. Shokrollahi, *Algebraic Complexity Theory*, vol. 315 of *Grundlehren der Mathematischen Wissenschaften*, Springer Verlag, Heidelberg, 1996.

[4] S. Gao and A. Shokrollahi, "Computing roots of polynomials over function fields of curves." In *Coding Theory and Cryptography*, David Joyner (Ed), 214–228, 1999, Springer-Verlag, New York.

[5] I. Gohberg and V. Olshevsky, "Fast state-space algorithms for matrix Nehari and Nehari-Takagi interpolation problems," *Integral Equations and Operator Theory* **20**, pp. 44–83, 1994.

[6] V. Guruswami and M. Sudan, "Improved decoding of Reed-Solomon and algebraic-geometric codes," in *Proceedings of the 39th IEEE Symposium on Foundations of Computer Science*, pp. 28–37, 1998.

[7] G. Heinig and K. Rost, *Algebraic Methods for Toeplitz-like matrices and operators*, vol. 13 of *Operator Theory*, Birkhäuser, Boston, 1984.

[8] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, North-Holland, 1988.

[9] T. Kailath and A. Sayed, "Displacement structure: Theory and applications," *SIAM Review* **37**, pp. 297–386, 1995.

[10] V. Olshevsky, "Pivoting for structured matrices with applications." Can be downloaded from `http://www.cs.gsu.edu/~matvro`, 1997.

[11] V. Olshevsky and V. Pan, "A superfast state-space algorithm for tangential Nevanlinna-Pick interpolation problem," in *Proceedings of the 39th IEEE Symposium on Foundations of Computer Science*, pp. 192–201, 1998.

[12] V. Olshevsky and A. Shokrollahi, "A displacement structure approach to list decoding of algebraic-geometric codes," in *Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, pp. 235–244, 1999.

[13] A. Shokrollahi and H. Wasserman, "Decoding algebraic-geometric codes beyond the error-correction bound," *Proceedings of the 30th Annual ACM Symposium on Theory of Computing* , pp. 241–248, 1998.

[14] M. Sudan, "Decoding of Reed-Solomon codes beyond the error-correction bound," *J. Compl.* **13**, pp. 180–193, 1997.

[15] L. Welch and E. Berlekamp, "Error correction for algebraic block codes." U.S. Patent 4,633,470, issued Dec. 30, 1986.

[16] S. Wicker and V. Bhargava, eds., *Reed-Solomon Codes and their Applications*, IEEE Press, 1994.

Algebraic Soft-Decision Decoding of Reed-Solomon Codes

Ralf Koetter
University of Illinois
Urbana, Illinois 61801
`koetter@csl.uiuc.edu`

Alexander Vardy
University of California
San Diego, CA 92093
`vardy@ece.uiuc.edu`

Abstract. A polynomial-time soft-decision decoding algorithm for Reed-Solomon codes is developed. The algorithm is algebraic in nature and builds upon the interpolation procedure proposed by Guruswami and Sudan for hard-decision decoding. Algebraic soft-decision decoding is achieved by means of converting the soft-decision reliability information into a set of interpolations points along with their multiplicities. The proposed conversion procedure is shown to be optimal for a certain probabilistic model. The resulting soft-decoding algorithm significantly outperforms both the Guruswami-Sudan decoding and the generalized minimum distance (GMD) decoding of Reed-Solomon codes, while maintaining a complexity that is polynomial in the length of the code.

1. Introduction

Reed-Solomon (RS) codes are among the most extensively used error-control codes, with applications ranging from magnetic recording, through satellite communications [2], to deep-space exploration. Since the discovery of these codes by Reed and Solomon [7] four decades ago, a steady stream of work has been devoted to their decoding. Today, very efficient hard-decision decoders for RS codes, using well-known algebraic decoding algorithms [11], are readily available. Such decoders have been implemented in hypersystolic VLSI to operate at sustained data rates in excess of 820 Mb/s.

An important problem in hard-decision decoding of Reed-Solomon codes is that of decoding beyond the error-correction radius (see e.g. [1] and references therein). A breakthrough in this area was recently achieved by Sudan [9, 4]. In the form presented in [4], the algorithm of Guruswami-Sudan corrects any fraction of $\tau \leqslant 1 - \sqrt{R}$ erroneous positions in a Reed-Solomon code of rate R. Thus the error-correction capability of the algorithm exceeds the minimum distance bound $(1 - R)/2$ for all rates in the interval $[0, 1)$.

Soft-decision decoding of Reed-Solomon codes is, however, an entirely different matter. Indeed, all the available *optimal* soft-decoding algorithms for RS codes, such as [10] and its various modifications, are non-algebraic and run in time that scales *exponentially* with the length of the code. This makes the use of such algorithms generally infeasible in practice. An alternative approach to the problem of efficient soft decoding was pioneered by Forney [3] and is known as generalized minimum distance (GMD) decoding. While the complexity of GMD decoding is moderate, and ultimately is of the same order as the complexity of hard-decision decoding [1], the gains that can be realized by GMD decoding are also moderate. In this context, Berlekamp et al. [2] state that "the major drawback with RS codes [· · ·] is that the present generation of decoders do not make full use of bit-based soft-decision information." In light of the ubiquity of Reed-Solomon codes, efficient soft-decision decoding of RS codes is one of the most important problems in coding theory and practice.

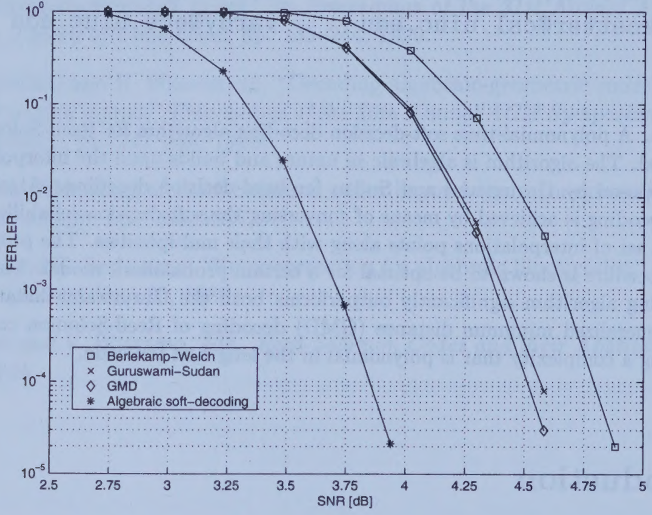


Figure 1. *Performance comparison for a simple concatenated coding scheme*

Here, a (256, 144, 113) Reed-Solomon code over GF(256) is concatenated with a binary (9, 8, 2) single-parity-check code, so that the overall rate is $^1/_2$. An additive white Gaussian noise (AWGN) channel is assumed, and the inner parity-check code is in all cases decoded with a maximum a posteriori likelihood decoding algorithm. The different curves correspond to the performance achieved by two hard-decision decoding algorithms and two soft-decision decoding algorithms. The two hard-decision algorithms are the conventional Berlekamp-Welch [11] decoding up to half the minimum distance and the algorithm by Guruswami and Sudan [4]. The two soft-decision algorithms are (a list decoding version of) Forney's GMD decoding [3] and the algebraic soft-decision list-decoding algorithm developed herein.

Our main goal in this paper is to present an efficient soft-decision decoding algorithm for Reed-Solomon codes. The algorithm is algebraic in nature and, for any desired level of performance (within a certain fundamental bound), its complexity is bounded by a polynomial in the codeword length. Our algorithm significantly outperforms both the Guruswami-Sudan decoding [4] and the GMD-based decoding methods. Figure 1 shows the performance of the three algorithms for a simple coding scheme. We note that this simple scheme was chosen solely to illustrate the performance gains that are possible with algebraic soft-decision decoding, rather than to achieve extreme coding gains.

Our algorithm is based on the algebraic interpolation techniques developed by Sudan in [4, 9]. To achieve soft-decision decoding, we translate the soft-decision reliability information provided by the channel into a set of algebraic constraints. More specifically, given the channel output vector $(y_1, y_2, \ldots, y_n)$ and the a posteriori transition probabilities $\Pr(c_i|y_i)$, we iteratively compute a set of interpolation points along with their multiplicities. We show that, at each step of the computation, this choice of interpolation points is optimal The complexity of the computation is $O(n^2 \log n)$, where the constant in $O(\cdot)$ depends on a certain design parameter. Notably, the algorithm of Guruswami-Sudan [9, 4] is based on algebraic interpolation and factorization techniques that can be implemented efficiently [6, 8, 12] in polynomial time. Our soft-decision decoding procedure inherits these properties of Guruswami-Sudan decoding. One of the most intriguing characteristics of our soft-decoding algorithm is a complexity/performance trade-off that can be chosen freely. In particular, the complexity can be easily adjusted to any required level of performance within a fundamental bound. Thus the coding gain provided by the Reed-Solomon code can be traded for complexity, in real-time, in any application.

2. Reed-Solomon codes and the Sudan algorithm

We first set up some of the notation that will be used throughout this work. Let $\mathbb{F}_q$ be the finite field with q elements. The ring of polynomials over $\mathbb{F}_q$ in a variable X is denoted $\mathbb{F}_q[X]$. Reed-Solomon codes are obtained by evaluating certain subspaces of $\mathbb{F}_q[X]$ in a set of points $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ which is a subset of $\mathbb{F}_q$. Specifically, the Reed-Solomon code $\mathbb{C}_q(n,k)$ of length n and dimension k is defined as follows:

$$\mathbb{C}_q(n,k) \stackrel{\text{def}}{=} \{(f(x_1),\ldots,f(x_n)) : x_1,\ldots,x_n \in \mathcal{D},\ f(X) \in \mathbb{F}_q[X],\ \deg f(X) < k\} \quad (1)$$

The point set $\mathcal{D}$ is usually taken as $\mathbb{F}_q$ or as $\mathbb{F}_q^*$, the set of all the nonzero elements of $\mathbb{F}_q$. Unless stated otherwise, we shall henceforth assume that $\mathcal{D} = \mathbb{F}_q^*$, so that $n = q-1$. The set of polynomials of degree less than k in $\mathbb{F}_q[X]$ is a linear space, which together with the linearity of the evaluation map (1) establishes that $\mathbb{C}_q(n,k)$ is a linear code.

Given an arbitrary vector $\underline{y} \in \mathbb{F}_q^n$, the *hard-decision decoding task* consists of finding the codeword $\underline{c} \in \mathbb{C}_q(n,k)$ such that the Hamming weight $\mathrm{wt}(\underline{e})$ of the error vector $\underline{e} = \underline{y} - \underline{c}$ is minimized. The Berlekamp-Welch algorithm [11] is a well-known algorithm that accomplishes this task, provided $\mathrm{wt}(\underline{e}) < d/2$. Generalizing upon Berlekamp-Welch [11], Sudan [9] and Guruswami-Sudan [4] derived a polynomial-time algorithm that achieves error correction substantially beyond half the minimum distance of the code. In the remainder of this section, we describe the essential elements of this algorithm.

Definition 1. *Let $\mathcal{A}(X,Y) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{i,j}X^iY^j$ be a bivariate polynomial over $\mathbb{F}_q$ and let w_X, w_Y be nonnegative real numbers. The (w_X, w_Y)-weighted degree of $\mathcal{A}(X,Y)$ is defined as the maximum over all numbers $iw_X + jw_Y$ such that $a_{i,j} \neq 0$.*

The $(1,1)$-weighted degree is simply the *degree* of a bivariate polynomial. The number of monomials of (w_X, w_Y)-weighted degree at most δ is denoted $N_{w_X,w_Y}(\delta)$. Thus

$$N_{w_X,w_Y}(\delta) \overset{\text{def}}{=} \left| \{X^i Y^j \ : \ i,j \geqslant 0 \text{ and } iw_X + jw_Y \leqslant \delta\} \right|$$

The following lemma provides a closed-form expression for $N_{w_X,w_Y}(\delta)$ for the case $w_X = 1$. Similar statements can be found in [4, 6, 8, 9].

Lemma 1.
$$N_{1,k}(\delta) = \left\lceil \frac{\delta+1}{k} \right\rceil \left(\delta - \frac{k}{2} \left\lfloor \frac{\delta}{k} \right\rfloor + 1 \right)$$

The lemma follows by a straightforward counting of monomials; for a detailed proof, see [4, Lemma 6]. The exact, but elaborate, expression in Lemma 1 can be converted into a simple lower bound as follows:

$$N_{1,k-1}(\delta) = \frac{(\delta+1)^2}{2(k-1)} + \frac{k-1}{2} \left(\left\lceil \frac{\delta+1}{k-1} \right\rceil - \left(\left\lceil \frac{\delta+1}{k-1} \right\rceil - \frac{\delta+1}{k-1} \right)^2 \right) > \frac{\delta^2}{2k} \tag{2}$$

This is, in fact, a special case of the more general lower bound $N_{w_X,w_Y}(\delta) > \delta^2/2w_X w_Y$. The latter bound can be easily proved using geometric arguments.

Definition 2. *A bivariate polynomial $\mathcal{A}(X,Y)$ is said to pass through a point (α,β) with multiplicity m if the shifted polynomial $\mathcal{A}(X+\alpha, Y+\beta)$ contains a monomial of degree m and does not contain a monomial of degree less than m. Equivalently, the point (α,β) is said to be a zero of multiplicity m of the polynomial $\mathcal{A}(X,Y)$.*

Using a well-known explicit relation (cf. [4]) between the coefficients of a bivariate polynomial $\mathcal{A}(X,Y)$ and the coefficients of the shifted polynomial, we find that Definition 2 imposes the following linear constraints

$$\sum_{i=k}^{\infty} \sum_{j=l}^{\infty} \binom{i}{k} \binom{j}{l} \alpha^{i-k} \beta^{j-l} a_{i,j} = 0 \qquad \forall\, k,l \geqslant 0 \text{ such that } k+l < m \tag{3}$$

on the coefficients $a_{i,j}$ of $\mathcal{A}(X,Y)$. Thus $\mathcal{A}(X,Y)$ passes through a given point with multiplicity at least m if and only if its coefficients $a_{i,j}$ satisfy the $^1\!/_2 m(m+1)$ constraints specified by (3). We are now ready to formulate the first step of the Sudan [4, 9] algorithm.

Interpolation step: *Given the set $\mathcal{P}$ of points in $\mathbb{F}_q \times \mathbb{F}_q$ and a positive integer m, compute the nontrivial bivariate polynomial $\mathcal{Q}_{\mathcal{P}}(X,Y)$ of minimal $(1,k-1)$-weighted degree that passes through all the points in $\mathcal{P}$ with multiplicity at least m.*

The number of linear constraints imposed on the coefficients of $\mathcal{Q}_{\mathcal{P}}(X,Y)$ by the interpolation step is $^1\!/_2\, nm(m+1)$. If the $(1,k-1)$ weighted degree of $\mathcal{Q}_{\mathcal{P}}(X,Y)$ is δ, then $\mathcal{Q}_{\mathcal{P}}(X,Y)$ may have up to $N_{1,k-1}(\delta)$ nonzero coefficients. These coefficients should be chosen so as to satisfy the $^1\!/_2\, nm(m+1)$ linear constraints of the type (3). This produces a system of

$\frac{1}{2}nm(m+1)$ linear equations (not all of them necessarily linearly independent) over $\mathbb{F}_q$ in $N_{1,k-1}(\delta)$ unknowns. It is clear that this system has a solution as long as

$$N_{1,k-1}(\delta) \; > \; \frac{nm(m+1)}{2} \tag{4}$$

For an efficient algorithm to solve such a system of linear equations and, hence, accomplish the interpolation step of the Sudan algorithm, we refer the reader to Nielsen and Høholdt [6] (but see also Roth and Ruckenstein [8]).

The idea of Sudan's algorithm [4, 9] is that, under certain constraints on the weight of the error vector, the relaxation of the classical interpolation problem to the problem of finding the polynomial $\mathcal{Q}_{\mathcal{P}}(X,Y)$ does not affect the solution. Hence we hope that $\mathcal{Q}_{\mathcal{P}}(X,Y)$ factors as $\Lambda(X,Y)(Y-f(X))$ and, reversing arguments, we can read-off a list of tentative decoding decisions as factors of $\mathcal{Q}_{\mathcal{P}}(X,Y)$ of type $Y-f(X)$. Thus the second (and last) step of the Sudan algorithm is as follows.

> **Factorization step:** *Given the bivariate polynomial $\mathcal{Q}_{\mathcal{P}}(X,Y)$, identify all the factors of $\mathcal{Q}_{\mathcal{P}}(X,Y)$ of type $Y-f(X)$ with $\deg f(X) < k$. The output of the algorithm is a list of the codewords that correspond to these factors.*

Factorization of $\mathcal{Q}_{\mathcal{P}}(X,Y)$ into factors of type $Y-f(X)$ with $\deg f(X) < k$ is a surprisingly simple task. Efficient algorithms to accomplish such factorization can be found in [6, 12], for example. The fundamental question is under which conditions can one guarantee that the correct decoding decision is found on the list produced by the factorization step. The answer to this question is given by the following theorem.

Theorem 2. *Suppose that a vector $\underline{y}$ and a positive integer m are given. Then the factorization step produces a list that contains all codewords of $\mathbb{C}_q(n,k)$ at distance less than*

$$t \;=\; n - \left\lfloor \frac{\delta}{m} \right\rfloor \;>\; \left\lfloor n\left(1-\sqrt{R\,\frac{m+1}{m}}\right)\right\rfloor \tag{5}$$

from $\underline{y}$, where δ is the smallest integer such that $N_{1,k-1}(\delta) > \frac{1}{2}nm(m+1)$ and $R = k/n$.

For a proof of (5), see [4, 6]. The inequality in (5) follows from (2). Here, we observe that Theorem 2 is a special case of Theorem 3, which we prove in the next section. Theorem 2 is the main result of Guruswami and Sudan in [4]. The theorem shows that as $m \to \infty$, the algorithm of [4] corrects any fraction of $\tau \leqslant 1-\sqrt{R}$ erroneous positions.

3. Algebraic soft-decision decoding

In many situations [2, 10], the decoder can be supplied with probabilistic reliability information concerning the received symbols. A decoding algorithm that utilizes such information is generally referred to as a *soft-decision* decoding algorithm. We now specify this notion precisely, in the context of the present paper. First, we define a ***memoryless channel***, or simply a ***channel***, as a collection of a finite input alphabet $\mathscr{X}$, an output alphabet $\mathscr{Y}$, and $|\mathscr{X}|$ functions

$$f(\cdot|x) : \mathscr{Y} \to [0,1] \qquad \text{for all } x \in \mathscr{X} \tag{6}$$

that are assumed to be known to the decoder. We think of channel input and output as random variables $\mathcal{X}$ and $\mathcal{Y}$, respectively, and assume that $\mathcal{X}$ is uniformly distributed over $\mathscr{X}$. If the channel is continuous (e.g. Gaussian), then $\mathcal{Y}$ is continuous and the $f(\cdot|x)$ are probability-density functions, while if the channel is discrete then $\mathcal{Y}$ is discrete and the $f(\cdot|x)$ are probability-mass functions. In either case, the decoder can easily compute the probability that $\alpha \in \mathscr{X}$ was transmitted given that $y \in \mathscr{Y}$ was observed, as follows

$$\Pr\Big(\mathcal{X}=\alpha \mid \mathcal{Y}=y\Big) = \frac{f(y|\alpha)\Pr(\mathcal{X}=\alpha)}{\sum_{x\in\mathscr{X}} f(y|x)\Pr(\mathcal{X}=x)} = \frac{f(y|\alpha)}{\sum_{x\in\mathscr{X}} f(y|x)} \tag{7}$$

where the second equality follows from the assumption that $\mathcal{X}$ is uniform. For Reed-Solomon codes, the input alphabet is always fixed to $\mathscr{X} = \mathbb{F}_q$. Henceforth, let $\alpha_1, \alpha_2, \ldots, \alpha_q$ be a fixed ordering of the elements of $\mathbb{F}_q$; this ordering will be implicitly assumed in the remainder of this paper. Given the vector $\underline{y} = (y_1, y_2, \ldots, y_n) \in \mathscr{Y}^n$ observed at the channel output, we compute

$$\pi_{i,j} \overset{\text{def}}{=} \Pr\Big(\mathcal{X}=\alpha_i \mid \mathcal{Y}=y_j\Big) \qquad \text{for } i = 1,2,\ldots,q \text{ and } j = 1,2,\ldots,n \tag{8}$$

according to the expression in (7). Let Π be the $q \times n$ matrix with entries $\pi_{i,j}$ defined in (8). We will refer to Π as the ***reliability matrix*** and assume that Π is the input to a soft-decision decoding algorithm. For notational convenience, we will sometimes write $\Pi(\alpha, j)$ to refer to the entry found in the j-th column of Π in the row indexed by $\alpha \in \mathbb{F}_q$.

The "soft" reliability information will enter the decoding process through the choice of interpolation points and their multiplicities. Given a reliability matrix Π, the key question is how to determine these points. An answer to this question is provided in the next section. In the remainder of this section, we characterize the proposed soft-decoding algorithm for a *given choice* of interpolation points and their multiplicities.

A convenient way to keep track of the interpolation points and their multiplicities is by means of a multiplicity matrix. A ***multiplicity matrix*** is a $q \times n$ matrix M with nonnegative integer entries $m_{i,j}$. The first step of our decoding algorithm consists of computing the multiplicity matrix M from the reliability matrix Π (cf. Algorithm A of the next section). The second step is the "soft" interpolation step, which may be expressed as follows.

Soft interpolation step: *Given the point set $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ and the multiplicity matrix $M = [m_{i,j}]$, compute the (nontrivial) bivariate polynomial $\mathcal{Q}_M(X,Y)$ of minimal $(1, k-1)$-weighted degree that has a zero of multiplicity at least $m_{i,j}$ at the point (x_j, α_i) for every i, j such that $m_{i,j} \neq 0$.*

The third and final step of our algorithm is the factorization step, which is identical to the factorization step of the Sudan algorithm, described in the previous section. (The soft-decision list-decoder may also include a post-processor that selects the most likely codeword from the list produced at the factorization step.)

It is easy to see from (3) and (4) that the computation of the polynomial $\mathcal{Q}_M(X,Y)$ is equivalent to solving a system of linear equations of the type (3). The complexity of this computation depends primarily on the *number* of equations in the system.

Definition 3. *Given a $q \times n$ matrix M with nonnegative integer entries $m_{i,j}$, we define the cost of M as follows*

$$\mathcal{C}(M) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{n} m_{i,j}(m_{i,j}+1)$$

As observed in (3), a given zero of multiplicity m imposes $1/2\, m(m+1)$ linear constraints on $\mathcal{Q}_M(X,Y)$. Thus $\mathcal{C}(M)$ is precisely the total number of linear equations. It follows that the complexity of computing $\mathcal{Q}_M(X,Y)$ is governed by the cost of the multiplicity matrix M. As in (4), we can always find a solution $\mathcal{Q}_M(X,Y)$ to the soft interpolation task if the $(1,k{-}1)$-weighted degree δ is large enough, namely if

$$N_{1,k-1}(\delta) > \mathcal{C}(M) \tag{9}$$

so that the number of degrees of freedom is greater than the number of linear constraints. The solution $\mathcal{Q}_M(X,Y)$ is said to *contain* a codeword $\underline{c} \in \mathbb{C}_q(n,k)$ if the factorization of $\mathcal{Q}_M(X,Y)$ contains a factor $Y - f(X)$ such that $\underline{c} = (f(x_1), f(x_2), \ldots, f(x_n))$. The fundamental question is under which conditions does the polynomial $\mathcal{Q}_M(X,Y)$ found in the soft interpolation step contain a given codeword $\underline{c} \in \mathbb{C}_q(n,k)$.

The answer to this question is given in Theorem 3 below. In order to state this theorem, we need some more notation. First, we define the function

$$\Delta_{w_X,w_Y}(\nu) \stackrel{\text{def}}{=} \min\{\delta \in \mathbb{Z} \,:\, N_{w_X,w_Y}(\delta) > \nu\} \tag{10}$$

Notice that $\Delta_{1,k-1}(\nu) \leqslant \sqrt{2k\nu}$ in view of (2). Next, given two $q \times n$ matrices A and B over the same field, we define the inner product

$$\langle A, B\rangle \stackrel{\text{def}}{=} \mathsf{trace}(AB^T) = \sum_{i=1}^{q}\sum_{j=1}^{n} a_{i,j}b_{i,j}$$

Finally, it will be convenient to think of the codewords of the Reed-Solomon code $\mathbb{C}_q(n,k)$ as $q \times n$ matrices over the reals. Specifically, any vector $\underline{v} = (v_1, v_2, \ldots, v_n)$ over $\mathbb{F}_q$ can be represented by the $q \times n$ real-valued matrix $[\underline{v}]$ defined as follows: $[\underline{v}]_{i,j} = 1$ if $v_j = \alpha_i$, and $[\underline{v}]_{i,j} = 0$ otherwise. Notice that, by construction, the matrix $[\underline{v}]$ contains precisely one nonzero entry in each column. With this notation, we have the following definition.

Definition 4. *The score of a vector $\underline{v} = (v_1, v_2, \ldots, v_n)$ over $\mathbb{F}_q$ with respect to a given multiplicity matrix M is defined as the inner product $\mathcal{S}_M(\underline{v}) = \langle M, [\underline{v}]\rangle$.*

The following theorem characterizes the set of codewords produced by our soft-decision list-decoding algorithm for a given multiplicity matrix. Notice that Theorem 2 follows as a special case of Theorem 3 for the multiplicity matrix $M = m\,[\underline{y}]$.

Theorem 3. *Let $\mathcal{C} = \mathcal{C}(M)$ denote the cost of a given multiplicity matrix M. Then the polynomial $\mathcal{Q}_M(X,Y)$ contains a codeword $\underline{c} \in \mathbb{C}_q(n,k)$ if the score of $\underline{c}$ is large enough compared to $\mathcal{C}$, namely if*

$$\mathcal{S}_M(\underline{c}) \; > \; \Delta_{1,k-1}(\mathcal{C}) \tag{11}$$

Proof. Let $\underline{c} = (c_1, c_2, \ldots, c_n)$ be a codeword of $\mathbb{C}_q(n,k)$, let $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ be the set of points used in the construction of $\mathbb{C}_q(n,k)$, and let $f(X)$ be the polynomial generating $\underline{c}$, so that $c_j = f(x_j)$ for $j = 1, 2, \ldots, n$. Given the bivariate polynomial $\mathcal{Q}_M(X,Y)$, we define the polynomial $g(X) \in \mathbb{F}_q[X]$ as follows

$$g(X) \; \stackrel{\text{def}}{=} \; \mathcal{Q}_M(X, f(X))$$

It would clearly suffice to prove that (11) implies that $g(X)$ is the all-zero polynomial, since then $\mathcal{Q}_M(X,Y)$ must be divisible by $Y - f(X)$. To prove that $g(X) \equiv 0$, we will show that $\deg g(X) \leqslant \Delta_{1,k-1}(\mathcal{C})$ and yet $g(X)$ has a factor of degree $\mathcal{S}_M(\underline{c})$. We write

$$\mathcal{S}_M(\underline{c}) \; = \; \langle M, [\underline{c}] \rangle \; = \; m_1 + m_2 + \cdots + m_n$$

where m_j denotes the entry found in the j-th column of M in the row indexed by c_j. Thus the polynomial $\mathcal{Q}_M(X,Y)$ passes through the point (x_j, c_j) with multiplicity at least m_j, for $j = 1, 2, \ldots, n$. We will make use of the following lemma.

Lemma 4. *Suppose that a bivariate polynomial $\mathcal{Q}(X,Y)$ passes through a point (α, β) in $\mathbb{F}_q \times \mathbb{F}_q$ with multiplicity at least m, and let $p(X)$ be any polynomial in $\mathbb{F}_q[X]$ such that $p(\alpha) = \beta$. Then the polynomial $\mathcal{Q}(X, p(X))$ is divisible by $(X - \alpha)^m$.*

This lemma is identical to Lemma 4 of [4], and we omit the proof. Since $f(x_j) = c_j$ for $j = 1, 2, \ldots, n$, it follows from Lemma 4 and the fact that $x_1, x_2, \ldots, x_n$ are all distinct that the polynomial $g(X) = \mathcal{Q}_M(X, f(X))$ is divisible by the product

$$(X - x_1)^{m_1} (X - x_2)^{m_2} \cdots (X - x_n)^{m_n}$$

whose degree is $\mathcal{S}_M(\underline{c})$. We conclude that either $\deg g(X) \geqslant \mathcal{S}_M(\underline{c})$ or $g(X) \equiv 0$. Since $\deg f(X) \leqslant k - 1$, it is obvious that the degree of $g(X) = \mathcal{Q}_M(X, f(X))$ cannot exceed the $(1, k-1)$-weighted degree of $\mathcal{Q}_M(X,Y)$. Yet it follows from (9) and (10) that the $(1, k-1)$-weighted degree of $\mathcal{Q}_M(X,Y)$ is precisely $\Delta_{1,k-1}(\mathcal{C})$. Thus if $g(X) \not\equiv 0$ then $\mathcal{S}_M(\underline{c}) \leqslant \deg g(X) \leqslant \Delta_{1,k-1}(\mathcal{C})$. Hence (11) implies that $g(X) \equiv 0$. ∎

4. From posterior probabilities to interpolation points

This section deals with the conversion of posterior probabilities derived from the channel output into a choice of interpolation points and their multiplicities. More specifically, given a reliability matrix Π, as defined in (8), we would like to compute the multiplicity matrix M that serves as the input to the soft interpolation step of our algorithm.

From Theorem 3 we see that the multiplicity matrix M should be chosen so as to maximize the score $\mathcal{S}_M(\underline{c})$ of the transmitted codeword $\underline{c} \in \mathbb{C}_q(n,k)$ for a given cost $\mathcal{C}(M)$. Let $\mathscr{M}_{q,n}$ denote the set of all $q \times n$ matrices with nonnegative integer entries $m_{i,j}$, and define

$$\mathscr{M}(\mathcal{C}) \overset{\text{def}}{=} \left\{ M \in \mathscr{M}_{q,n} \ : \ \frac{1}{2}\sum_{i=1}^{q}\sum_{j=1}^{n} m_{i,j}(m_{i,j}{+}1) = \mathcal{C} \right\}$$

Thus $\mathscr{M}(\mathcal{C})$ is the finite set of all matrices in $\mathscr{M}_{q,n}$ whose cost is equal to $\mathcal{C}$, and our problem is that of finding the optimal matrix in $\mathscr{M}(\mathcal{C})$. We shall see that while this problem is apparently intractable for an arbitrary $\mathcal{C}$, for certain specific values of $\mathcal{C}$ the problem has a simple polynomial-time solution. This will generally suffice for our purposes.

Next, we elaborate upon what *optimality* means in our context. As explained above, we would like to choose $M \in \mathscr{M}(\mathcal{C})$ so as to maximize the score of the transmitted codeword $\underline{c} \in \mathbb{C}_q(n,k)$. However, the transmitted codeword itself is obviously unknown to the decoder; only some stochastic information about $\underline{c}$ is available through the observation of the channel output $(y_1, y_2, \ldots, y_n) \in \mathscr{Y}^n$ and the knowledge of the channel transition probabilities $\Pr(\mathcal{X} = \alpha \,|\, \mathcal{Y} = y)$. In fact, as far as the decoder is concerned, the transmitted codeword may be thought of as a random vector, which we denote by $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n)$. For a given multiplicity matrix M, the score of the transmitted codeword is a function of $\boldsymbol{\mathcal{X}}$ given by $\mathcal{S}_M(\boldsymbol{\mathcal{X}}) = \langle M, [\boldsymbol{\mathcal{X}}] \rangle$. Thus $\mathcal{S}_M(\boldsymbol{\mathcal{X}})$ is a random variable, and a reasonable optimization criterion is to choose $M \in \mathscr{M}(\mathcal{C})$ so as to maximize the expected value of $\mathcal{S}_M(\boldsymbol{\mathcal{X}})$. Specifically, we define the *expected score* with respect to a probability distribution $P(\cdot)$ on $\boldsymbol{\mathcal{X}}$ as follows

$$\mathsf{E}_P\{\mathcal{S}_M(\boldsymbol{\mathcal{X}})\} \overset{\text{def}}{=} \sum_{\underline{x} \in \mathscr{X}^n} \mathcal{S}_M(\underline{x}) P(\underline{x}) = \sum_{\underline{x} \in \mathbb{F}_q^n} \sum_{j=1}^{n} M(x_j, j) P(\underline{x}) \tag{12}$$

where $M(x_j, j)$ denotes the entry found in the j-th column of M in the row indexed by x_j. It remains to specify $P(\cdot)$. For this purpose, we adopt the product distribution determined by the channel output $(y_1, y_2, \ldots, y_n) \in \mathscr{Y}^n$, namely

$$P(x_1, x_2, \ldots, x_n) \overset{\text{def}}{=} \prod_{j=1}^{n} \Pr\Big(\mathcal{X}_j = x_j \mid \mathcal{Y}_j = y_j\Big) = \prod_{j=1}^{n} \Pi(x_j, j) \tag{13}$$

where Π is the reliability matrix defined in (8). It is easy to see that this would be the *a posteriori* distribution of $\boldsymbol{\mathcal{X}}$ given the channel observations, if the *a priori* distribution of $\boldsymbol{\mathcal{X}}$ were uniform over the space $\mathbb{F}_q^n$. The remainder of this section is concerned with the computation of the matrix $M(\Pi, \mathcal{C})$ defined as follows

$$M(\Pi, \mathcal{C}) \overset{\text{def}}{=} \operatorname{argmax}_{M \in \mathscr{M}(\mathcal{C})} \mathsf{E}_P\{\mathcal{S}_M(\boldsymbol{\mathcal{X}})\} \tag{14}$$

where the expectation is taken with respect to the probability distribution $P(\cdot)$ in (13). We start with the following lemma, which we give without a proof.

Lemma 5. *The expected score with respect to the probability distribution in* (13) *is equal to the inner product of the multiplicity matrix and the reliability matrix, namely*

$$\mathsf{E}_P\{\mathcal{S}_M(\boldsymbol{\mathcal{X}})\} = \langle M, \Pi \rangle$$

We will construct $M(\Pi, C)$ iteratively, starting with the all-zero matrix and increasing one of the entries in the matrix at each iteration. Referring to Lemma 5, we see that increasing $m_{i,j}$ from 0 to 1 increases the expected score by $\pi_{i,j}$ while increasing the cost by 1. If we require that $\mathcal{Q}_M(X,Y)$ passes through the same point again — that is, increase $m_{i,j}$ from 1 to 2 — then the expected score again grows by $\pi_{i,j}$, but now we have to "pay" two additional linear constraints. In general, increasing $m_{i,j}$ from a to $a+1$ always increases the expected score by $\pi_{i,j}$ while introducing $a+1$ additional constraints of type (3).

These observations suggest a greedy iterative procedure for the allocation of interpolation points, which turns out to be optimal. We formalize this procedure as Algorithm A below.

Algorithm A

Input: Reliability matrix Π and a positive integer s, indicating the total number of interpolation points.

Output: Multiplicity matrix M.

Initialization step: Set $\Pi^* := \Pi$ and $M :=$ `all-zero matrix`.

Iteration step: Find the position (i,j) of the largest entry $\pi^*_{i,j}$ in Π^*, and set

$$\begin{aligned} \pi^*_{i,j} &:= \frac{\pi_{i,j}}{m_{i,j}+2} \\ m_{i,j} &:= m_{i,j}+1 \\ s &:= s-1 \end{aligned}$$

Control step: If $s = 0$, return M; otherwise go to the iteration step.

Let $\mathcal{M}(\Pi, s)$ denote the multiplicity matrix produced by Algorithm A for a given reliability matrix Π and a given number of interpolation points s (counted with multiplicities). The following theorem, presented here without a proof, shows that this matrix is optimal.

Theorem 6. *The matrix $\mathcal{M}(\Pi, s)$ maximizes the expected score among all matrices in $\mathscr{M}_{q,n}$ with the same cost. That is, if C is the cost of $\mathcal{M}(\Pi, s)$, then*

$$\mathcal{M}(\Pi, s) \;=\; \operatorname{argmax}_{M \in \mathscr{M}(C)} \langle M, \Pi \rangle$$

Although Algorithm A produces an optimal multiplicity matrix $\mathcal{M}(\Pi, s)$ for an arbitrary number of interpolation points s, it cannot be used to solve the optimization problem (14) for an arbitrary value of the cost C. The algorithm computes a solution to (14) only for those costs that are expressible as the total length of the s highest rectangles in $\mathfrak{B}$ for some s. In other words, if $M(\Pi,1), M(\Pi,2), \ldots$ is the infinite sequence of matrices defined by (14) for $C = 1, 2, 3, \ldots$, then $\mathcal{M}(\Pi,1), \mathcal{M}(\Pi,2), \ldots$ is a subsequence of this sequence. On the other hand, it is shown in [5] that this is the best one could hope for.

References

[1] E.R. Berlekamp, "Bounded distance+1 soft-decision Reed-Solomon decoding," *IEEE Trans. Inform. Theory*, vol. 42, pp. 704–721, May 1996.

[2] E.R. Berlekamp, R.E. Peile, and S.P. Pope, "The application of error control to communications," *IEEE Commun. Mag.*, vol. 25, pp. 44–57, January 1987.

[3] G.D. Forney, Jr., "Generalized minimum distance decoding," *IEEE Trans. Inform. Theory*, vol. 12, pp. 125–131, April 1966.

[4] V. Guruswami and M. Sudan, "Improved decoding of Reed-Solomon and algebraic-geometric codes," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1755–1764, 1999.

[5] R. Koetter and A. Vardy "Algebraic soft-decision decoding of Reed-Solomon codes," preprint, June 2000.

[6] R.R. Nielsen and T. Høholdt, "Decoding Reed-Solomon codes beyond half the minimum distance," preprint, 1998; see also `http://www.student.dtu.dk/ p938546`.

[7] I.S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *J. Soc. Indust. Appl. Math.*, vol. 8, pp. 300–304, June 1960.

[8] R.M. Roth and G. Ruckenstein, "Efficient decoding of Reed-Solomon codes beyond half the minimum distance", *IEEE Trans. Inform. Theory*, vol. 46, pp. 246–258, 2000.

[9] M. Sudan, "Decoding of Reed-Solomon codes beyond the error correction bound," *J. Complexity*, vol. 12, pp. 180–193, 1997.

[10] A. Vardy and Y. Be'ery, "Bit-level soft-decision decoding of Reed-Solomon codes," *IEEE Trans. Commun.*, vol. 39, pp. 440–445, March 1991.

[11] L.R. Welch and E.R. Berlekamp, *Error correction for algebraic block codes*, U.S. Patent No. 4,633,470, issued December 30, 1986.

[12] X-W. Wu and P.H. Siegel, "Efficient list decoding of algebraic-geometric codes beyond the error correction bound," *IEEE Trans. Inform. Theory*, submitted May 2000.

On the (Non-)Convergence of the EM Algorithm for Discrete Parameter Estimation*

Predrag Spasojević
WINLAB, Dept. of ECE
Rutgers University
Piscataway, NJ 08854
spasojev@winlab.rutgers.edu

Costas N. Georghiades
Electrical Engineering Department
Texas A&M University
College Station, TX 77843

Abstract

It is well known that the EM algorithm converges to a likelihood stationary point in convex parameter spaces when a set of quite general continuity and regularity conditions holds. On the other hand, its non-convergence to the maximum-likelihood solution in the case of discrete parameter spaces has been observed.

In this work, it is shown that a number of fixed points of the discrete EM algorithm may exist within a non-convergence ball centered at a fixed point of a companion unconstrained EM algorithm. The radius of the non-convergence ball is equal to the ratio of the minimum distance of the discrete parameter set and the largest convergence step of the companion unconstrained EM algorithm. In the case of sequence estimation in Gaussian interference with orthogonal signaling a reduced non-convergence radius can be related to the interference spectrum. By forcing a rank reduction in the interference model one may significantly increase the probability of convergence.

1 Introduction

Even though stability and monotonicity of the EM algorithm have long been valued for both discrete and convex parameter spaces, method's wider application to estimation in discrete parameter spaces has been hampered by the fact that its convergence to even a

---

*This work was supported in part by National Science Foundation grant NCR-9314221.

local maximum has not been proved. It is well known that the EM algorithm converges to a likelihood stationary point in convex parameter spaces when a set of quite general continuity and regularity conditions holds (see [1] and [2]). On the other hand, its non-convergence to the maximum-likelihood solution in the case of discrete parameter spaces has been observed (see, e.g., [3]).

In this work, it is shown that a number of fixed points of the discrete EM algorithm may exist within a non-convergence ball centered at a fixed point of a companion unconstrained EM algorithm. The radius of the non-convergence ball is equal to the ratio of the minimum distance of the discrete parameter set and the largest convergence step of the companion unconstrained EM algorithm. In the case of sequence estimation in Gaussian interference with orthogonal signaling, a reduced non-convergence radius can be related to the interference spectrum. By forcing a rank reduction in the interference model one may significantly increase the probability of convergence.

The companion unconstrained EM algorithm is introduced in the second section. Third section introduces the non-convergence ball for the general case. Fourth section describes the reduced non-convergence ball and the convergence induced using rank-reduction methods for the interference problem.

2 EM-Based Discrete and Its Companion Unconstrained Parameter Estimation

Let $\mathbf{r}$ be a sufficient statistic for estimation of the N-dimensional parameter $\mathbf{a}$.

The discrete parameter ML estimation

$$\hat{\mathbf{a}}_d = \arg\max_{\mathbf{a}\in\mathbf{\Omega_d}} p(\mathbf{r}|\mathbf{a}) \tag{1}$$

requires maximization of the likelihood function over a discrete parameter set $\mathbf{\Omega}_d$. In the following, it is assumed that, additionally, a companion unconstrained maximization problem

$$\max_{\mathbf{a}\in\mathcal{C}^{\mathrm{N}}} p(\mathbf{r}|\mathbf{a}) \tag{2}$$

is well defined. Here, $\mathcal{C}^N$ the space of all complex N by 1 vectors.

Given any current estimate $\mathbf{a}^k$, the next iterate of the discrete EM algorithm for solution of (1) is

$$\mathbf{a}_d^{k+1} = \arg\max_{\mathbf{a}\in\mathbf{\Omega_d}} \ Q(\mathbf{a}|\mathbf{a}^k). \tag{3}$$

The next iterate of its companion unconstrained EM algorithm for solution of (2) is

$$\mathbf{a}_c^{k+1} = \arg\max_{\mathbf{a}\in\mathcal{C}^N} Q(\mathbf{a}|\mathbf{a}^k). \tag{4}$$

The discrete EM mapping is defined as the map $\mathbf{M}_d^e : \mathbf{a}^k \to \mathbf{a}_d^{k+1}$, for $\mathbf{a}_d^{k+1} \in \mathbf{\Omega}_d$. On the other hand, the companion unconstrained EM mapping is the map $\mathbf{M}_c^e : \mathbf{a}^k \to \mathbf{a}_c^{k+1}$, for $\mathbf{a}_c^{k+1} \in \mathcal{C}^N$. The Jacobian of the unconstrained EM mapping $\mathbf{M}_c(\mathbf{a})$ is denoted with $\mathbf{J}_c(\mathbf{a})$ and the Hessian of its corresponding objective function $Q(\mathbf{a}|\mathbf{a}^k)$ is denoted with $\mathbf{H}^Q(\mathbf{a})$.

3 The Non-Convergence Ball of the Discrete EM Algorithm with a Quadratic Objective Function

An approximate analysis based on the Taylor series expansion of the companion unconstrained EM mapping is employed for derivation of a set of fixed points of the discrete EM algorithm which are not, in the general case, local maximizers of the log-likelihood function. Note that, for unconstrained parameter estimation, under continuity and weak regularity conditions, stationary points of the EM algorithm have to be stationary points of the log-likelihood function (see, e.g., [4]). On the other hand, any fixed point can be a convergence point of the discrete EM algorithm even though the likelihood of such a point can be lower than the likelihood of a neighboring discrete point.

Computationally "inexpensive" maximization of the objective function $Q(\mathbf{a}|\mathbf{a}^k)$ can be obtained for some special structures. The particular case of interest is the quadratic form

$$Q(\mathbf{a}|\mathbf{a}^k) \;=\; -\frac{1}{2}\|\mathbf{a} - \mathbf{a}_c^{k+1}\|^2 \tag{5}$$

for which the discrete maximization in the *M − Step* can be obtained on a parameter-by-parameter basis. Equation (5) implies that $\mathbf{a}_d^{k+1}$ is the discrete point closest to the unconstrained maximizer of $Q(\mathbf{a}|\mathbf{a}^k)$, $\mathbf{a}_c^{k+1}$. Note that $\mathbf{a}_c^{k+1}$ is (implicitly) a function of the previous iterate $\mathbf{a}^k$. As a special case of (5), $\mathbf{a}_d^{k+1} = \text{sign}[\mathbf{a}_c^{k+1}]$ for QPSK signaling, i.e., for $a_{d,i} \overset{\triangle}{=} [\mathbf{a}_d]_i \in \{\pm 1 \pm j\}$. Clearly, the objective function (5) is not only quadratic but, additionally, its Hessian is $\mathbf{H}^Q(\mathbf{a}) = -\mathbf{I}$. Nevertheless, the arguments given here can be generalized to a non-identity Hessian matrix.

Let d_{min} denote half the minimum Euclidean distance between any two discrete points from $\mathbf{\Omega}_d$, i.e.,

$$d_{min} = \frac{1}{2}\min_{\mathbf{a}_{d,1},\mathbf{a}_{d,2}\in\mathbf{\Omega}_d} \|\mathbf{a}_{d,1} - \mathbf{a}_{d,2}\|. \tag{6}$$

$d_{min} = 1$ for uncoded binary antipodal signaling. In this case, the discrete points are vertices of a N-dimensional cube centered at the origin.

From (3) and (5) we have that

$$\mathbf{a}_d^{k+1} = \arg\min_{\mathbf{a}\in\mathbf{\Omega}_d} \|\mathbf{a} - \mathbf{a}_c^{k+1}\|. \tag{7}$$

Let the previous iterate $\mathbf{a}^k$ in (5) (and, implicitly, in (7)) be a discrete point $\mathbf{a}_d^k \in \mathbf{\Omega}_d$ such that

$$\|\mathbf{a}_d^k - \mathbf{a}_c^{k+1}\| < d_{min}. \tag{8}$$

Now, (6), (7), and (8) imply that $\mathbf{a}_d^{k+1} = \mathbf{a}_d^k$. In other words, $\mathbf{a}_d^k$ for which (8) holds is a stationary point of the discrete EM algorithm.

Let $\hat{\mathbf{a}}_c$ be a fixed point of the companion unconstrained EM algorithm (4). It is assumed that, in a neighborhood $U(\hat{\mathbf{a}}_c) \subseteq \mathcal{C}^N$ of $\hat{\mathbf{a}}_c$, $\mathbf{M}_c(\mathbf{a})$ can be approximated sufficiently closely by the first two terms of its Taylor series expansion. That is,

$$\mathbf{M}_c(\mathbf{a}^k) \approx \mathbf{M}_c(\hat{\mathbf{a}}_c) + \mathbf{J}_c(\hat{\mathbf{a}}_c)(\mathbf{a}^k - \hat{\mathbf{a}}_c), \tag{9}$$

or, equivalently,

$$\mathbf{a}_c^{k+1} \approx \hat{\mathbf{a}}_c + \mathbf{J}_c(\hat{\mathbf{a}}_c)[\mathbf{a}^k - \hat{\mathbf{a}}_c]. \tag{10}$$

for any $\mathbf{a}^k \in U(\hat{\mathbf{a}}_c)$. Simple manipulations lead to the following approximate identity:

$$\mathbf{a}_c^{k+1} - \mathbf{a}^k \approx [\mathbf{I} - \mathbf{J}_c(\hat{\mathbf{a}}_c)](\hat{\mathbf{a}}_c - \mathbf{a}^k). \tag{11}$$

Equations (8) and (11) imply (approximately) that all discrete points $\mathbf{a}^k = \mathbf{a}_d^k \in U(\hat{\mathbf{a}}_c)$ such that

$$\|[\mathbf{I} - \mathbf{J}_c(\hat{\mathbf{a}}_c)](\hat{\mathbf{a}}_c - \mathbf{a}_d^k)\| < d_{min} \tag{12}$$

are stationary points of the discrete EM algorithm. The non-convergence condition can be simplified by observing that the inequality (12) holds if

$$\|\hat{\mathbf{a}}_c - \mathbf{a}_d^k\| < \frac{d_{min}}{1 - \hat{\lambda}_{min}}, \tag{13}$$

where $\hat{\lambda}_{min} \triangleq \lambda_{min}[\mathbf{J}_c(\hat{\mathbf{a}}_c)]$ is the minimum eigenvalue of the Jacobian matrix $\mathbf{J}_c(\hat{\mathbf{a}}_c)$. Inequality (13) follows from the fact that the eigenvalues of $\mathbf{J}_c(\hat{\mathbf{a}}_c)$ are non-negative and smaller than one (see Lemma in [5]). It defines the *nonconvergence ball* of radius

$$r^{nc} = d_{min}/(1 - \hat{\lambda}_{min})$$

centered at a fixed point of the companion unconstrained EM algorithm $\hat{\mathbf{a}}_c$. All the discrete points contained in this ball (and in the neighborhood $U(\hat{\mathbf{a}}_c)$) are stationary points of the discrete EM algorithm.

Parameter $\hat{\lambda}_{min}$ is, in the general case, a function of the received signal. As shown by Meng and Rubin [6] (see also [4]), $\mathbf{J}_c(\hat{\mathbf{a}}_c)$ and, consequently, its eigenvalues can be estimated in a reliable manner using the unconstrained EM algorithm. Furthermore, when the complete data of the EM algorithm is chosen in such a way that $\mathbf{H}^Q(\hat{\mathbf{a}}_c) = -\mathbf{I}$ as in (5), then matrices $\mathbf{J}_c(\hat{\mathbf{a}}_c)$ and the Hessian of the log-likelihood function, $\mathbf{H}^l(\hat{\mathbf{a}}_c)$, have the same eigenvectors. Their eigenvalues have the following relationship: $1 - \lambda_i[\mathbf{J}_c(\hat{\mathbf{a}}_c)] = \lambda_i[-\mathbf{H}^l(\hat{\mathbf{a}}_c)]$, for all i. Thus,

$$r^{nc} = \frac{d_{min}}{\lambda_{max}[-\mathbf{H}^l(\hat{\mathbf{a}}_c)]}.$$

$(1 - \hat{\lambda}_{min})$ is the largest eigenvalue of the matrix $\mathbf{I} - \mathbf{J}_c(\hat{\mathbf{a}}_c)$ often referred to as the iteration matrix in the optimization literature (see, e.g., [4]) since it determines the convergence iteration steps in the neighborhood $U(\hat{\mathbf{a}}_c)$ as per equation (11). Its smallest eigenvalue is a typical measure of convergence speed of the unconstrained EM algorithm since it measures the convergence step along the slowest direction. Its largest eigenvalue determines the convergence step along the fastest direction. Thus, we can say that the non-convergence radius is determined by (inversely proportional to) the largest convergence step of the companion unconstrained EM algorithm.

Figure 1 illustrates most of the concepts described in this section. It depicts the non-convergence ball of radius $r^{nc} = 1$, the minimum distance $d_{min} = 1/2$, and the largest convergence step $(1 - \hat{\lambda}_{min}) = 1/2$. We can observe that in a neighborhood of the maximizer $\hat{\mathbf{a}}_c$ the iteration matrix $\mathbf{I} - \mathbf{J}_c(\hat{\mathbf{a}}_c)$ of the unconstrained EM mapping maps the difference $(\hat{\mathbf{a}}_c - \mathbf{a}^k)$ into the difference $\mathbf{a}_c^{k+1} - \mathbf{a}^k$. Since the latter difference is smaller then the (half) minimum distance d_{min}, a quantizer from (7) or, equivalently, the M-step of the discrete EM algorithm, quantizes $\mathbf{a}_c^{k+1}$ back to $\mathbf{a}^k$. Thus, the algorithm can not move to a more likely discrete point which is closer to $\hat{\mathbf{a}}_c$.

An important implication of (13) is that $\hat{\lambda}_{min} = 0$ is sufficient for the non-convergence ball to hold at most one discrete stationary point. Sequence estimation in the presence of interference when $\hat{\lambda}_{min} = 0$ or when this identity can be forced using rank-reduction principles has been analyzed in [7], [3], and presented in part in the next section.

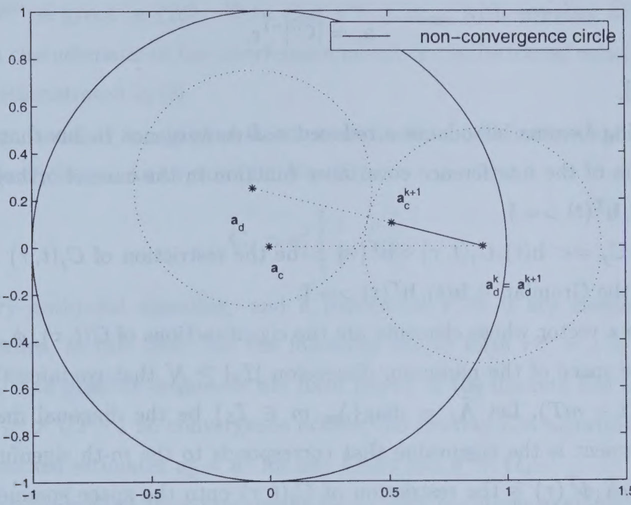


Figure 1: The non-convergence ball of the discrete EM algorithm.

4 Detection in Gaussian Interference

The general unconstrained-time model of the received modulated signal in a Gaussian interference channel is (see [3])

$$r(t) \;=\; \sum_{n=0}^{N-1} a_n h(t-nT) + j(t) + n(t) = \mathbf{a}^T\mathbf{h}(t) + j(t) + n(t). \tag{14}$$

N is the number of transmitted symbols; $h(t)$ is the signal pulse; T is the symbol interval; a_n is the n-th transmitted symbol; $j(t)$ is a circularly symmetric complex zero-mean Gaussian process with a self-adjoint covariance function $C_j(t,\tau)$; $n(t)$ is a circularly symmetric complex zero-mean white Gaussian WSS process with a covariance function $N_0\delta(t-\tau)$; $\mathbf{h}(t) = [h(t), h(t-T), \ldots, h(t-(N-1)T)]^T$ is the N by 1 pulse vector which incorporates all symbol pulses. The symbol sequence vector has the same form: $\mathbf{a} = [a_1, \ldots, a_N]^T$.

We define $\mathbf{C}^i = N_0 < \mathbf{h}(t), C^{-1}(t,\tau) \diamond \mathbf{h}^T(\tau) >$ as the normalized restriction of $C^{-1}(t,\tau)$ onto the signal space spanned by the elements of $\mathbf{h}(t)$. $C^{-1}(t,\tau)$, the inverse total covariance function, is a solution of the following integral equation: $< C^{-1}(t,u), C(u,\tau) >=< C^{-1}(t,u), C_j(u,\tau) > +N_0C^{-1}(t,\tau) = \delta(t-\tau)$. $\mathbf{r} = N_0 < \mathbf{h}(t), C^{-1}(t,\tau) \diamond r(\tau) >$ is the sufficient statistic for the problem.

The center of the non-convergence ball for this problem has the closed form

$$\hat{\mathbf{a}}_c = [\mathbf{C}^i]^{-1}\mathbf{r},$$

as shown in [3].

The following Lemma introduces a reduced non-convergence radius that can be related to the spectrum of the interference covariance function in the case of orthogonal signaling, i.e., for $< \mathbf{h}(t), \mathbf{h}^T(t) >= \mathbf{I}$.

Lemma: Let $\mathbf{C}_j =< \mathbf{h}(t), C_j(t,\tau) \diamond \mathbf{h}^T(\tau) >$ be the restriction of $C_j(t,\tau)$ onto the signal space, and let the Gramian $< \mathbf{h}(t), \mathbf{h}^T(t) >= \mathbf{I}$.

Let $\boldsymbol{\phi}(t)$ be a vector whose elements are the eigenfunctions of $C(t,\tau)$, $\phi_m(t)$ for $m \in \mathcal{I}_S$, which span the space of the minimum dimension $|\mathcal{I}_S| \geq \mathcal{N}$ that contains the signal space spanned by $h(t-mT)$. Let $\tilde{\boldsymbol{\Lambda}}_j = \text{diag}\{\lambda_m, m \in \mathcal{I}_S\}$ be the diagonal matrix whose m-th diagonal element is the eigenvalue that corresponds to the m-th eigenfunction in $\boldsymbol{\phi}(t)$. $\tilde{C}_j(t,\tau) = \boldsymbol{\phi}(t)\tilde{\boldsymbol{\Lambda}}_j\boldsymbol{\phi}^T(\tau)$ is the restriction of $C_j(t,\tau)$ onto the space spanned by $\phi_m(t)$ for $m \in \mathcal{I}_S$.

Then, all discrete points contained in a ball centered at $\hat{\mathbf{a}}_c$ with a radius

$$\tilde{r}^{nc} = \left[1 + \frac{\tilde{\lambda}_{min}}{N_0}\right] d_{min}, \tag{15}$$

where $\tilde{\lambda}_{min} = \lambda_{min}[\tilde{C}_j(t,\tau)]$ are stationary points of the EM algorithm.

This ball is contained in the non-convergence ball defined in Section 3, that is, $\tilde{r}^{nc} \leq r^{nc}$.

Furthermore, if $|\mathcal{I}_S| = \mathcal{N}$, i.e., if $\text{span}\{h(t-nT), n \in [0, N-1]\} = \text{span}\{\phi_i(t), i \in \mathcal{I}_S\}$ then

$$\begin{aligned} \tilde{r}^{nc} &= \left[1 + \frac{\lambda_{min}(\mathbf{C}_j)}{N_0}\right] d_{min} & (16) \\ &= r^{nc}. & (17) \end{aligned}$$

The proof is given in [3]. No approximations are used since in the derivation of the reduced non-convergence ball since (9) holds without approximation for all $\mathbf{a}^k \in U(\hat{\mathbf{a}}_c) = \mathcal{C}^N$.

Note that $\tilde{\lambda}_{min} = 0$ iff the interference null-space contains some signal energy.

Corollary: The Lemma implies that the reduced non-convergence ball contains at most one discrete point when the interference null-space contains some signal energy. One can easily show that this will be the case iff $\lambda_{min}(\mathbf{C}_j) = 0$. In this case, $\tilde{r}^{nc} = r^{nc} = d_{min}$.

The proof is given in [3].

In the first example, for simplicity, it is assumed that $|\mathcal{I}_{\mathcal{S}}| = \mathcal{N}$. The non-convergence radius $r^{nc} = \tilde{r}^{nc}$ is given in (16). Note that $r^{nc} \geq d_{min}$ with equality iff $\lambda_{min}(\mathbf{C}_j) = 0$ which can be a characteristic of the interference model or can be forced using rank reduction principles as demonstrated in [3].

On the other hand, if $\lambda_{min}(\mathbf{C}_j) \neq 0$, one can choose a sufficiently small N_0 so that r^{nc} becomes arbitrarily large. As an example $N = 2$,

$$\mathbf{C}_j = \sigma_j^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

uncoded binary antipodal signaling, and a realization $\mathbf{r} \approx \mathbf{0}$ are assumed. The non-convergence radius, in this case, has the following simple form $r^{nc} = 1 + \frac{\sigma_j^2}{N_0}(1 - \rho)$. For $r^{nc} > \sqrt{N} = \sqrt{2}$ all possible sequences are fixed points of the discrete EM algorithm. E.g., for $\rho = 0$ and $\frac{\sigma_j^2}{N_0} > \sqrt{2} - 1$ no convergence occurs, the discrete EM algorithm generates the sequence of identical estimates $\mathbf{a}_d^k = \mathbf{a}^0$ for any k and any $\mathbf{a}^0 \in \mathbf{\Omega}_d$.

Next, we provide symbol error-rate results for a communications example where $|\mathcal{I}_{\mathcal{S}}| > \mathcal{N}$ holds. Figure 2 demonstrates that, in the case of an infinite-dimensional spectrum, one can significantly increase the convergence probability of the EM algorithm using a rank-reduction approach. Two second order Gaussian processes uniformly distributed in frequency within the signal spectra are used to model interference. The second-order interference process has a rational spectrum which is non-zero for all frequencies ω. The optimal receiver requires all available dimensionality for its complete representation. The EM algorithm obtained by modeling the interference with a number of largest eigenfunctions of its covariance $C_j(t, \tau)$ is compared to the performance of the EM algorithm that uses the full rank model. It can be observed that the reduced-rank EM algorithm has little degradation, whereas, the full-rank EM algorithm does not converge and has a catastrophic performance for large SNR.

5 Conclusion

It is shown that a number of fixed points of the discrete EM algorithm may exist within a non-convergence ball centered at a fixed point of a companion unconstrained EM algorithm. The radius of the non-convergence ball is equal to the ratio of the minimum distance of the discrete parameter set and the largest convergence step of the companion unconstrained EM algorithm. In the case of sequence estimation in Gaussian interference with orthogonal signaling a reduced non-convergence radius can be related to the interference spectrum.

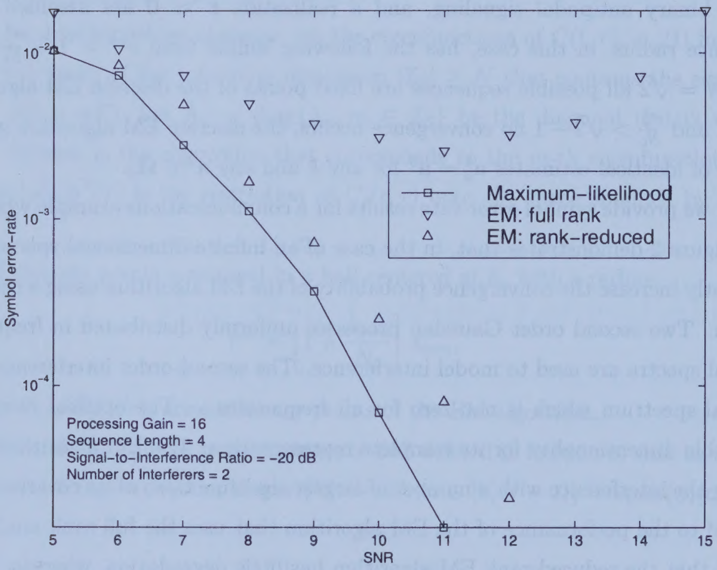


Figure 2: Comparison of ML, reduced-rank EM and full-rank EM receivers for 2 second-order Gaussian interferers.

By forcing a rank reduction in the interference model one may significantly increase the probability of convergence.

References

[1] A. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorthm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–17, 1977.

[2] C. Wu, "On the convergence properties of the EM algorthm," *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[3] P. Spasojević, *Sequence and channel estimation for channels with memory.* Dissertation: Texas A&M University, College Station, TX, December, 1999.

[4] G. J. Machlachlan and T. Krishnan, *The EM Algorithm and Extensions.* New York: John Wiley and Sons, 1997.

[5] X. L. Meng and D. B. Rubin, "On the global and componentwise rates of convergence of the EM algorthm," *Linear Algebra and Its Applications*, vol. 199, pp. 413–425, 1994.

[6] X. L. Meng and D. B. Rubin, "Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm," *Journal of the American Statistical Association*, vol. 86, pp. 899–909, 1991.

[7] C. N. Georghiades and P. Spasojević, "Maximum-likelihood detection in the presence of interference," submitted to *IEEE Transactions on Communications*, 1999.

An alternating minimization approach to image formation in the presence of a known function*

Ryan J. Murphy, Joseph A. O'Sullivan, and Donald L. Snyder
Washington University
St. Louis, Missouri 63130

Abstract

The problem of image formation when part of the image is known up to some parameters is addressed. This consists of estimating the parameters associated with the known part and also estimating the remainder of the image. An alternating minimization algorithm for this class of problems is then applied to computerized tomographic imaging in the presence of known rigid objects; specifically, for the radiation treatment known as *brachytherapy*. The resulting algorithms are more computationally demanding than standard filtered back projection, but provide improved images and estimates of the pose of the known object.

It is well-known that the computerized tomographic (CT) scanning of high-density metal objects, such as surgical clips and metal rods, in soft tissue produces severe streaking artifacts in the images reconstructed using conventional filtered back projection. Modern CT scanners send a fan-shaped beam of X-rays through an object (often the human body) and an array of detectors opposite the X-ray source measures the attenuated source-intensity. High-density metal in the presence of tissue, however, can attenuate the beam significantly. The result of this when using conventional filtered back projection techniques is that we see many alternating bright and dark streaks emanating from the metal in the resulting image.

It is, therefore, of great interest to determine this unknown function (the image) better. This can be viewed as a linear, deterministic, deblurring problem. We are given a nonnegative function $\{m(y), y \in Y\}$ (the sensor data, or *sinogram*), and a nonnegative kernel $\{h(y|x), y \in Y, x \in X\}$, which acts as a point spread function characterizing the sensor. For our problem, X and Y are discrete index sets. Elements of X index spatial locations, and elements of Y index source-detector pairs. We search for a function, $\{c(x), x \in X\}$ (the image), such that $m(y) \approx \sum_{x \in X} h(y|x)c(x)$, $(y \in Y)$ is a good approximation. We can think of $c(x)$ as a linear combination of an unknown function, $c_b(x)$, and a known function, $c_a(x : \theta)$, parameterized by some unknown parameter, θ. This known function could incorporate any kind of *a priori* information about $c(x)$; in our example, this could be knowledge of the metal structures. We use Csiszár's I-divergence as a discrepancy measure for determining when this approximation of $c(x)$ is good enough [1]. The I-divergence between two positive vectors $\mathbf{p}$ and $\mathbf{r}$ is defined as follows: $I(\mathbf{p}\|\mathbf{r}) \equiv \sum_{j=1}^{n}(p_j \log(p_j/r_j) - p_j + r_j)$.

---

*This work was supported in part by Washington University and in part by the National Cancer Institute of the National Institutes of Health under research grant R01CA75371, (J. Williamson, P. I.)

The image we would like to reconstruct is of the form $c(x) = c_a(x : \theta) + c_b(x)$, where $c_a(x : \theta)$ represents some known function with unknown parameters, θ, and $c_b(x)$ is an unknown image. The model for the data is

$$b(y : \theta) = \sum_{x \in X} h(y|x)c(x) = \mathbf{Hc}. \tag{1}$$

The problem is then to compute $[\mathbf{c}_b, \theta] = \operatorname{argmin}_{\mathbf{c}_b,\theta} I(\mathbf{m}||\mathbf{Hc})$, where $\mathbf{m}$ is the measured data. This I-divergence can be written as a minimization of an expanded I-divergence, $I(\mathbf{m}||\mathbf{Hc}) = \min_{\mathbf{P} \in \mathcal{P}} I(\mathbf{Pm}||\mathbf{Hc})$, where $\mathcal{P}$ is a set of probability matrices [2].

The alternating minimization algorithm alternates between updating the estimate for $c(x)$ with $\mathbf{P}$ held fixed and updating the estimate for $\mathbf{P}$ with $c(x)$ held fixed. If $\hat{c}^{(k)}(x)$ is the estimate for $c(x)$ at iteration k, then the next estimate for $\mathbf{P}$ is $\hat{\mathbf{P}}^{(k)}$ with entries

$$\hat{p}^{(k)}(x|y) = \frac{h(y|x)\hat{c}^{(k)}(x)}{\sum_{x' \in X} h(y|x')\hat{c}^{(k)}(x')}. \tag{2}$$

We propose a generalized alternating minimization algorithm for this problem, which updates $c(x) = c_a(x : \theta) + c_b(x)$ to an improved estimate (lowering the value of the I-divergence), but does not find the global minimum at each step. In the odd iterations, the estimate for θ is updated (i.e., we minimize over θ), while in the even iterations we minimize over $c_b(x)$, repeating until adequate convergence is attained.

We apply these methods to improve the quality of images used in *brachytherapy*, a type of cancer treatment that involves implanting a metal applicator containing sealed, radioactive sources in or next to a malignant tumor. High-density metal in the applicator (used for shielding benign areas) can introduce artifacts into the CT scans. The unknown parameters, $\theta \in \Theta$, represent the applicator's position and orientation. This applicator image is then added to the reconstructed image of the surrounding tissue, $c_b(x)$, to form the estimate, $\hat{c}(x)$. We alternately improve the guesses for the pose and the tissue image until acceptable convergence is reached.

We synthetically create the measured data, $m(y)$, by forward projecting a two dimensional slice of the body with the applicator superimposed, and then ran 600 iterations of the alternating minimzation algorithm. We have shown significantly less streaking, both in number and magnitude, than a similar experiment using filtered back projection. Additionally, this alternating minimization approach provides knowledge of the applicator's pose—in this instance within 0.25 mm [3].

References

[1] I. Csiszár, "Why least squares and maximum entropy?—an axiomatic approach to inverse problems," *Ann. Stat.*, vol. 19, pp. 2033–2066, Dec. 1991.

[2] J. A. O'Sullivan, "Alternating minimization algorithms: from Blahut-Arimoto to expectation-maximization," in: *Codes, Curves, and Signals—Common Threads in Communications* (A. Vardy, ed.), Kluwer Academic Publishers, pp. 173–192, 1998.

[3] D. L. Snyder, J. A. O'Sullivan, B. R. Whiting, R. J. Murphy, J. Benac, J. A. Cataldo, and J. F. Williamson, "Deblurring subject to nonnegativity constraints when known functions are present, with application to object-constrained computerized tomography," submitted for publication June 2000.

Spatial Spectrum Estimation for Time-Varying Arrays Using the EM Algorithm*

David W. Rieken[1,†], Aaron D. Lanterman[2], and Daniel R. Fuhrmann[1]

[1]Electronic Systems and Signals Research Laboratory
Dept. of Electrical Engineering, Campus Box 1127
Washington University in St. Louis
St. Louis, MO 63130

[2]Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
1308 W. Main Street
Urbana, IL 61801

October 5, 2000

Abstract

We examine the problem of estimating the spatial distribution of incident radiation using sensor arrays in which the relative element locations change during the observation interval. To this end a maximum-likelihood estimator is derived by making several realistic assumption about the statistics of the incident radiation. This estimator represents an extension of the structured covariance estimation problem for sensor arrays to the case of time-varying array manifolds. The estimator is applied to simulated data specific to sonar arrays.

1 Introduction

The advantage that arrays of sensors have over single sensors, in addition to the increase in gain, is the ability to discriminate direction without moving the sensor. This capability manifests itself in several forms, one of which is to estimate the distribution of incident power as a function of direction of arrival. The range of angles considered depends on, among other things, the geometry of the array. A linear array, for example, will only be capable of estimating radiated power density on S^1 (S^n is the n-sphere). While there are many algorithms for performing this task, they break down when the relative position of the array elements is time-varying. We consider the problem of spatial spectrum estimation using an array whose shape changes significantly during the data collection interval.

It is easy to visualize scenarios in which the array will change shape while data is being collected. Perhaps the most notorious example is that of underwater hydrophone arrays, in which the rope-like structures of the (nominally) linear array in its fluid environment

---

*This research is supported in part by a contract from MIT Lincoln Laboratory

†`rieken@essrl.wustl.edu`

makes it susceptible to time-variations. Underwater currents cause the array to change shape quickly in much the same way as a breeze causes a flag to change its shape. Were one to apply existing spectrum estimation algorithms to data collected over intervals in which the sonar array changed shape, a decrease in performance should be expected. Other examples of time-varying arrays can be found in airborne radar arrays in which the elements are spread over some non-rigid portion of the aircraft, as well as in formations of single-sensor platforms.

While their has been a wealth of work for the non-time-varying case, there is little for the time-varying case. A Maximum Likelihood Estimator (MLE) for estimating direction of arrival using time-varying arrays can be found in [1]. Here it is assumed that the number of sources and the geometry of the array at each sample are known and the corresponding direction of arrival is derived. Fast algorithms for doing the same are presented in [2] and [3]. We will build on the this work by deriving an estimator for the power density on the far-field sphere using many of the same assumptions. We will follow the derivation of a similar algorithm for spatial spectrum estimation using a stationary array which may be found in [4]. Also, the resulting algorithm is similar to the radio astronomy-specific estimator in [5]. We proceed by first deriving a log-likelihood function. An estimator which finds a local maxima of the function is then arrived at using the iterative EM algorithm. Finally, the algorithm will be tested on simulated sonar array data.

2 Likelihood Function

Let us assume that there is an N-element array from which M data vectors are extracted. We begin by modeling the array output, $\mathbf{x}_m$, as

$$\mathbf{x}_m = \int s(\Theta; t_m)\mathbf{a}(\Theta; t_m)\, d\Theta + \mathbf{w}_m \quad 0 \le m < M \tag{1}$$

where $\mathbf{a}(\Theta; t_m) \in \mathbf{C}^{\mathbf{N}}$ is the array response to a signal originating in direction Θ at time t_m (commonly and henceforth referred to as the *steering vector*), $s(\Theta; t_m) \in \mathbf{C}$ is the contribution of the differential emitter at $\Theta \in \mathbf{S}^2$ to the signal at the array at time t_m, and $\mathbf{w}_m \in \mathbf{C}^N$ is receiver noise. We will assume that the steering vector is known for all t_m. We can discretize (1) by dividing $\mathbf{S}^2$ into K discrete components by which we arrive at

$$\mathbf{x}_m = \mathbf{A}(t_m)\mathbf{s}_m + \mathbf{w}_m \tag{2}$$

where

$$\mathbf{A}(t_m) = \begin{bmatrix}\mathbf{a}(\Theta_1; t_m) & \mathbf{a}(\Theta_2; t_m) & \cdots & \mathbf{a}(\Theta_K; t_m)\end{bmatrix}$$

and

$$\mathbf{s}_m = \begin{bmatrix} s(\Theta_1; t_m)\, \partial\Theta_1 \\ s(\Theta_2; t_m)\, \partial\Theta_2 \\ \vdots \\ s(\Theta_K; t_m)\, \partial\Theta_K \end{bmatrix}.$$

Since the continuous distribution of emitters has been approximated by a discrete distribution we will shorten the notation by henceforth using $s_k(t_m)$ to denote $s(\Theta_k; t_m)\partial\Theta_k$.

We will assume that emitters at different spatial locations are uncorrelated. Therefore $\mathbf{s}_m$ is a complex normal random variable with mean zero and covariance

$$P = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \sigma_K^2 \end{bmatrix}. \tag{3}$$

Notice that σ_k^2 is the power of the kth discretized emitter and the quantity to be estimated. We will also assume that the sample spacing is sufficiently large so that $\mathrm{E}\left[\mathbf{s}_m\mathbf{s}_n^H\right] = \delta_{mn}$. Finally, we will assume that $\mathbf{w}_m$ is zero-mean, complex normal and that it is uncorrelated over time: $\mathrm{E}[\mathbf{w}_m\mathbf{w}_n^H] = N_0\mathbf{I}\delta_{nm}$. We will also assume that N_0 is known.

As $\mathbf{x}_m$ is the sum of normal random variables, it is a normal random variable. We calculate its covariance as

$$\begin{aligned} \mathbf{R}_m &= \mathrm{E}\left[\mathbf{x}_m\mathbf{x}_m^H\right] \\ &= \mathbf{A}(t_m)\mathbf{P}\mathbf{A}(t_m) + N_0\mathbf{I}. \end{aligned} \tag{4}$$

Note that for time-invariant arrays $\frac{\partial}{\partial t}\mathbf{a}(\Theta; t) = 0$ for all t, so that $\mathbf{R}_0 = \mathbf{R}_1 = \cdots = \mathbf{R}_{M-1}$. Covariance estimation under these circumstances is addressed in [6] and for block-Toeplitz and Toeplitz-block-Toeplitz matrix structure in [7]. Spectral estimation is considered for the radar imaging case in [8], for general array processing in [4], and for radio astronomy in [5]. What follows should be viewed as a generalization of these estimators.

The probability density function of $\mathbf{x}_m$ may be written

$$f_{\mathbf{x}_m}(\mathbf{x}_m) = \pi^{-N}\left|\mathbf{R}_m\right|^{-1}\exp\left(-\mathbf{x}_m^H\mathbf{R}_m^{-1}\mathbf{x}_m\right). \tag{5}$$

Since successive samples of the data are uncorrelated the joint density function is

$$\begin{aligned} f_{\mathbf{x}_0\cdots\mathbf{x}_{M-1}}(\mathbf{x}_0, \cdots, \mathbf{x}_{M-1}) &= \prod_{m=0}^{M-1} f_{\mathbf{x}_m}(\mathbf{x}_m) \\ &= \pi^{-NM}\left(\left|\mathbf{R}_0\right|^{-1}\left|\mathbf{R}_1\right|^{-1}\cdots\left|\mathbf{R}_{M-1}\right|^{-1}\right)\exp\left(-\sum_{m=0}^{M-1}\mathbf{x}_m^H\mathbf{R}_m^{-1}\mathbf{x}_m\right). \end{aligned} \tag{6}$$

The log-likelihood function can then be stated:

$$L(\boldsymbol{\sigma}^2) = \sum_{m=0}^{M-1}\left(\ln\left|\mathbf{R}_m^{-1}\right| - \mathbf{x}_m^H\mathbf{R}_m^{-1}\mathbf{x}_m\right) \tag{7}$$

We have expressed the parameters to be estimated collectively as a vector so that the ith element of $\boldsymbol{\sigma}^2$ is σ_i^2. A closed form solution to the likelihood function is not immediately obvious. We therefore seek an iterative solution.

3 EM Algorithm

The Expectation-Maximization (henceforth EM) algorithm is presented as an iterative technique for MLE in [9]. We summarize the algorithm here. The data space is first divided into an incomplete data space which is typically the raw data output, and a complete data space which must satisfy the following properties: 1) there is a deterministic many-to-one mapping of the complete data space to the incomplete data space, and 2) the MLE problem is easy given the complete data. Selection of the complete data space is less a science and more an art. Log-likelihood functions are calculated on each data space. Each iteration consists of an expectation step and a maximization step. In the expectation step, the following function is evaluated:

$$Q\left(\mathbf{x}|\hat{\mathbf{x}}^{(l)}\right) \equiv \mathrm{E}\left[L_{CD}\left(\mathbf{x}\right)|\text{incomplete data}, \hat{\mathbf{x}}^{(l)}\right] \tag{8}$$

Here, $\mathbf{x}$ is the parameter to be estimated and $L_{CD}\left(\mathbf{x}\right)$ is the log-likelihood function on the complete data space. $\hat{\mathbf{x}}^{(l)}$ is the lth iterate in the estimate of the complete data. The maximization step is the selection of $\mathbf{x}$ such that $Q\left(\mathbf{x}|\hat{\mathbf{x}}^{(l)}\right)$ is maximized. Under moderate conditions successive iterations will produce $\mathbf{x}$ which maximizes the likelihood function.

We follow the derivation of [4] for the stationary array case. We begin by selecting as our incomplete data space the data vectors $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{M-1}$. The complete data space we specify as the data vectors $\mathbf{s}_0$, $\mathbf{s}_1$, $\cdots$, $\mathbf{s}_{M-1}$ and $\mathbf{w}_0$, $\mathbf{w}_1$, $\cdots$, $\mathbf{w}_{M-1}$. Assuming independence, the joint probability density function for the complete data is

$$\begin{aligned} f_{\mathbf{s}_0\cdots\mathbf{s}_{M-1},\mathbf{w}_0\cdots\mathbf{w}_{M-1}}\left(\mathbf{s}_0, \cdots, \mathbf{s}_{M-1}, \mathbf{w}_0, \cdots, \mathbf{w}_{M-1}\right) &= \prod_{m=0}^{M-1} \pi^{-N}|\mathbf{P}|^{-1}\exp\left[-\mathbf{s}_m^H\mathbf{P}^{-1}\mathbf{s}_m\right] \\ &\cdot \prod_{m=0}^{M-1}\frac{1}{N_0}\pi^{-N}\exp\left[-\frac{\mathbf{w}_m^H\mathbf{w}_m}{N_0}\right] \end{aligned} \tag{9}$$

from which we obtain the complete data log-likelihood function:

$$L_{CD}\left(\sigma^2\right) = -M\ln|\mathbf{P}| + \sum_{m=0}^{M-1}\left[-\mathbf{s}_m^H\mathbf{P}\mathbf{s}_m - \frac{\mathbf{w}_m^H\mathbf{w}_m}{N_0}\right]. \tag{10}$$

3.1 Expectation Step

Using (8) we get

$$Q\left(\sigma^2|\hat{\sigma^2}^{(l)}\right) = -M\ln|\mathbf{P}| - \sum_{m=0}^{M-1}\mathrm{E}\left[\mathbf{s}_m^H\mathbf{P}^{-1}\mathbf{s}_m|\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{M-1}, \hat{\sigma^2}^{(l)}\right]. \tag{11}$$

If we let $s_{km} = s\left(\Theta_k; t_m\right)$ then we can simplify the expression on the right by noting that

$$\mathrm{E}\left[\mathbf{s}_m^H\mathbf{P}^{-1}\mathbf{s}_m|\mathbf{x}_m, \hat{\sigma^2}^{(l)}\right] = \sum_{k=1}^{K}\mathrm{E}\left[\frac{|s_k(t_m)|^2}{\sigma_k^2}|\mathbf{x}_m, \hat{\sigma_k^2}^{(l)}\right]. \tag{12}$$

It is therefore necessary in evaluating the above that we first determine the probability density function of $s_k(t_m)|\mathbf{x}_m$. It is well-known that given zero-mean normal random vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}|\mathbf{y}$ is also normal with mean given by $\mathbf{R_{xy}}\mathbf{R_{yy}}^{-1}\mathbf{y}$ and covariance $\mathbf{R_{xx}} - \mathbf{R_{xy}}\mathbf{R_{yy}}^{-1}\mathbf{R_{yx}}$. The cross-covariance of $s_k(t_m)$ and $\mathbf{x}_m$ is

$$\begin{aligned} \mathrm{E}\left[s_k(t_m)\mathbf{x}_m^H\right] &= \mathrm{E}\left[s_k(t_m)\left(\sum_{j=1}^{K}s_j(t_m)\mathbf{a}(\Theta_j; t_m) + \mathbf{w}_m\right)\right] \\ &= \sigma_k^2\mathbf{a}^H\left(\Theta_k; t_m\right) \end{aligned} \tag{13}$$

from which we can deduce the conditional mean and covariance of s_{km} as

$$\operatorname{var}\left[s_k(t_m)|\mathbf{x}_m, \hat{\sigma_k^2}^{(l)}\right] = \hat{\sigma_k^2}^{(l)} - \left(\hat{\sigma_k^2}^{(l)}\right)^2 \mathbf{a}^H(\Theta_k;t_m)\mathbf{R}_m^{-1}\mathbf{a}(\Theta_k;t_m) \tag{14}$$

$$\mathrm{E}\left[s_k(t_m)|\mathbf{x}_m, \hat{\sigma_k^2}^{(l)}\right] = \hat{\sigma_k^2}^{(l)} \mathbf{a}^H(\Theta_k;t_m)\mathbf{R}_m^{-1}\mathbf{x}_m. \tag{15}$$

The expression on the right of (12) can now be determined:

$$\begin{aligned}\mathrm{E}\left[\frac{|s_k(t_m)|^{2\,(l)}}{\sigma_k^2}\ |\mathbf{x}_m, \hat{\sigma_k^2}^{(l)}\right] = \frac{1}{\sigma_k^2}\Bigg(\hat{\sigma_k^2}^{(l)} - \left(\hat{\sigma_k^2}^{(l)}\right)^2 \mathbf{a}^H(\Theta_k;t_m)\hat{\mathbf{R}}_m^{-1}\mathbf{a}(\Theta_k;t_m) \\ + \left(\hat{\sigma_k^2}^{(l)}\right)^2 \mathbf{a}^H(\Theta_k;t_m)\hat{\mathbf{R}}_m^{-1}\mathbf{x}_m\mathbf{x}_m^H\hat{\mathbf{R}}_m^{-1}\mathbf{a}(\Theta_k;t_m)\Bigg).\end{aligned} \tag{16}$$

This completes the derivation of the expectation function.

3.2 Maximization Step

Since $\mathbf{P}$ is a diagonal matrix we can re-write the expectation function in terms of the spatial power parameters, $\hat{\sigma_k^2}$. Substituting (3) and (12) into (11) yields

$$Q\left(\sigma^2|\hat{\sigma^2}^{(l)}\right) = \sum_{k=1}^{K}\left(-M\ln\sigma_k^2 - \sum_{m=0}^{M-1}\frac{1}{\sigma_k^2}\mathrm{E}\left[\|s_k(t_m)\|^2|\mathbf{x}_m, \hat{\sigma_k^2}^{(l)}\right]\right). \tag{17}$$

We need to find the σ^2 which maximizes this expression. The kth component of the gradient of the expectation function with respect to σ^2 is easily determined:

$$\frac{\partial Q}{\partial\sigma_k^2} = -\frac{M}{\sigma_k^2} + \frac{1}{\sigma_k^4}\sum_{m=0}^{M-1}\left(\mathrm{E}\left[\|s_k(t_m)\|^2|\mathbf{x}_m, \hat{\sigma_k^2}^{(l)}\right]\right). \tag{18}$$

Setting the above expression equal to zero at each k and solving for σ_k^2 gives the desired result. Combining this with the result from (16) we get

$$\begin{aligned}\hat{\sigma_k^2}^{(l+1)} = \hat{\sigma_k^2}^{(l)} + \frac{\left(\hat{\sigma_k^2}^{(l)}\right)^2}{M}\sum_{m=0}^{M-1}\Bigg[&\mathbf{a}^H(\Theta_k;t_m)\hat{\mathbf{R}}_m^{-1}\mathbf{x}_m\mathbf{x}_m^H\hat{\mathbf{R}}_m^{-1}\mathbf{a}(\Theta_k;t_m) \\ &- \mathbf{a}^H(\Theta_k;t_m)\hat{\mathbf{R}}_m^{-1}\mathbf{a}(\Theta_k;t_m)\Bigg].\end{aligned} \tag{19}$$

This is the update equation to be used at each iteration of the algorithm. It can be shown that in addition to successive estimates having increasing log-likelihood, they will also be positive. This is a fortunate consequence since negative power estimates would be paradoxical.

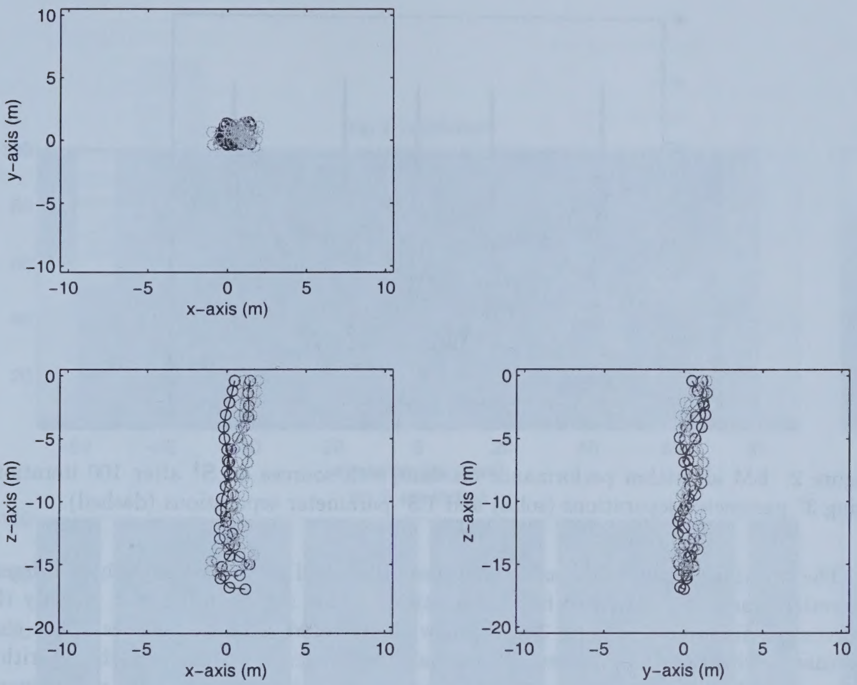


Figure 1: Three views of the time-varying arrays used in both simulations at four different times.

4 Simulation Results

We simulate performance of the EM algorithm using a nominally linear underwater array of omni-directional hydrophones. The array is analogous to towed sonar arrays found on many modern naval vessels with the exception that it is mounted to a buoy rather than a self-propelled platform. The fluidity of the water causes the array to flutter in the current in much the same way as wind causes flags to move. Orthographic projections of the simulated array at several different times are shown in figure 1. Here the $N = 20$ acoustic sensors are indicated as circles and are connected to each other with separations of 1 m. Time snapshots are discriminated by color. Care has been taken to model the drag, buoyancy, and gravity on each element in a moderately changing sea water environment so as to realistically replicate an underwater vertical sonar array. The array geometry is simulated at $M = 128$ consecutive, evenly spaced intervals. The time separation has been chosen so that it is long enough to decorrelate source signals and to affect significant change in the array shape from sample to sample.

So as to reduce the computational complexity of the scenario $K = 5$ discrete sources have been placed in the far-field of the x-z plane. The sources are located at −75°, −30°, 0°, 30° and 75° elevation. Each source has a SNR of 30dB at any single element of the array. We need only evaluate the likelihood function on the circle which lies in the x-z plane. Note that if it was not known that the sources were confined to this subspace the likelihood function would have to be evaluated on $\mathbf{S^2}$.

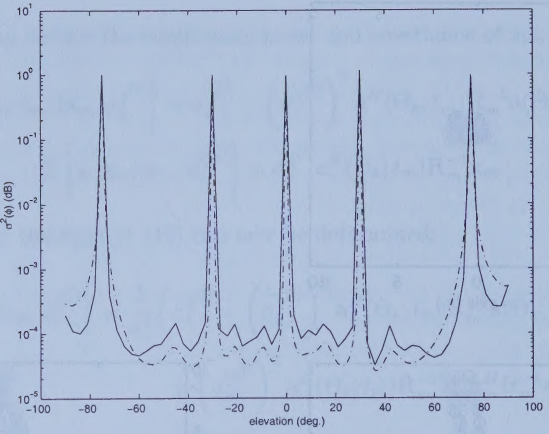
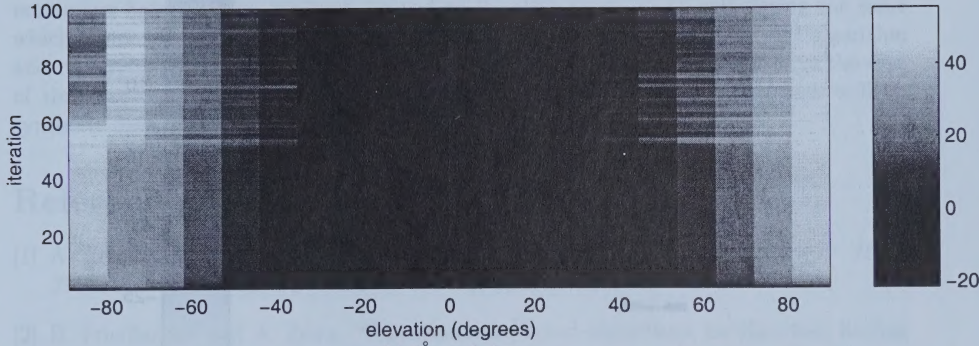


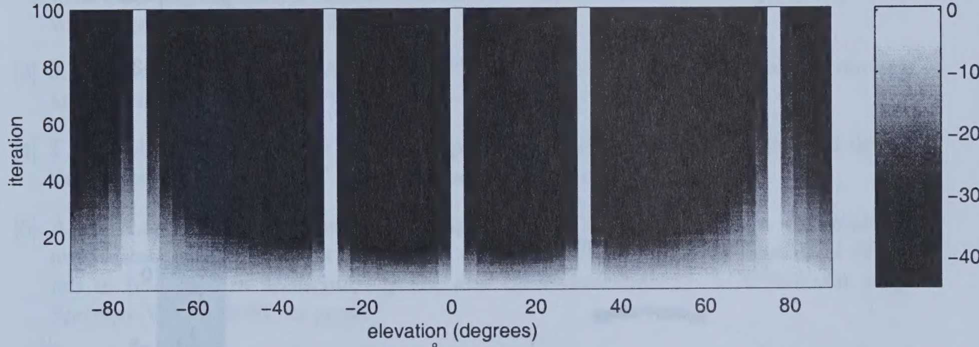
Figure 2: EM algorithm performance on data with sources on $\mathbf{S}^1$ after 100 iterations using 3° parameter separations (solid) and 1.5° parameter separations (dashed).

The iterative EM algorithm as derived above is applied to the data with three different discretizations of $\mathbf{S}^1$. The two finest discretizations are shown in figure 2. Clearly the sources are discernible. Figure 3 shows how the iterated estimate behaves. Note that the case in which we have discretized the evaluation space at 9° intervals, the algorithm does not produce convergent iterates. This indicates a minimum resolution requirement. At 3° and 1.5° discretizations the iterates appear to converge. For the 3° experiment the estimate converges to within 30dB after 30 iterations near 0° elevation and roughly 70 iterations near the poles. For the 1.5° experiment the estimate reaches this level of accuraccy after 50 iterations on the equator and more than 100 at the poles. The reader should note that despite the motion of the array, its extent still lies primarily along the z-axis and its performance will be much better at the equator than at the poles. These experimental results suggest that there may be an optimal discretization of the evaluation space. Should the discretization be too coarse, the power density estimates may not converge. Should it be too fine, the estimate will converge, but do so slowly.

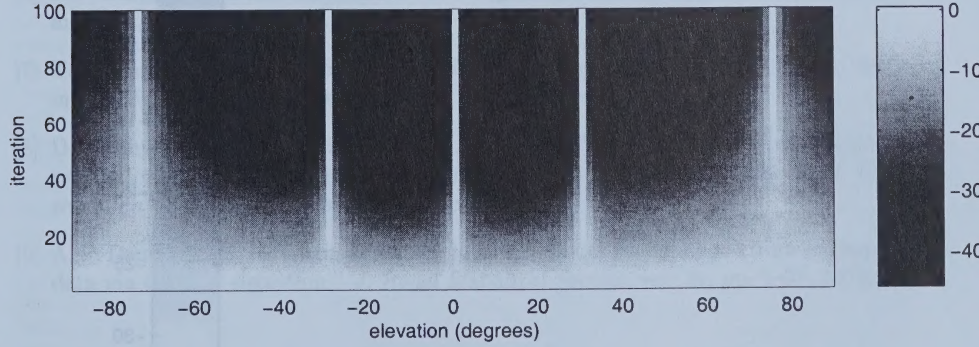
Finally, the fact that the array traces out a "synthetic" array which is not a linear array suggests that it is possible to resolve sources in azimuth. For stationary linear arrays, this is not the case. To demonstrate this capability, a second data set in which the 5 emitters are separated in azimuth as well as elevation was synthesized. The emitter locations are at (azimuth,elevation)=$(-180°, -75°)$, $(0°, -30°)$, $(180°, 0°)$, $(-90°, 30°)$, $(90°, 75°)$. The EM algorithm is applied to the data set and the results presented in figure 4. The estimated parameters were set up on a grid in azimuth and elevation so there are more pixels per solid angle near the poles than at the equator. Note that the noise estimate at the poles is not as good as the estimate at the equator. However, the source locations are easily picked out, even at the poles. Were this algorithm to be used in an actual sonar system, the spectrum estimation step may be followed by using the direction of arrival estimator in [2]. The main utility of the spectrum estimation in this case is in determining that there are 5 sources and their approximate directions from the sensor array.



(a) 9° resolution



(b) 3° resolution



(c) 1.5° resolution

Figure 3: Plot of EM algorithm performance on data with sources on $\mathbf{S}^1$ demonstrating the relationship between the number of parameters and rate of convergence. Results are in dB.

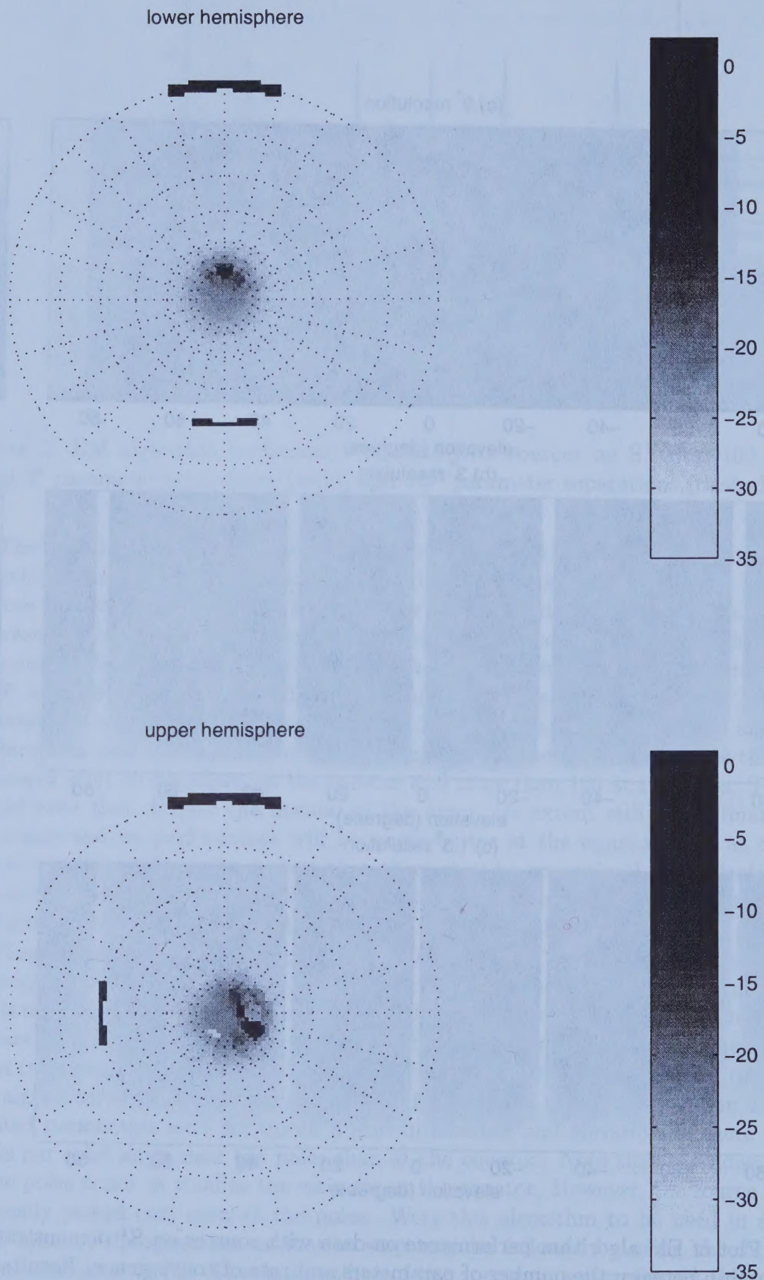


Figure 4: EM algorithm performance on data with sources on $\mathbf{S^2}$ after 300 iterations. Each plot is by stereo-graphic projection with the outer ring corresponding to the equator and concentric circles separated by 15° in elevation. Results are in dB.

5 Conclusions

We have examined the problem of spatial power spectrum estimation for moving arrays and derived a likelihood function for the power density. Closed-form solutions for the maximizer being elusive, we have derived an iterative approach to calculating the same which builds on the work of others. The algorithm is an application of the EM algorithm and represents an extension of the structured covariance estimation problem to the case of time-varying arrays. Simulation results show that the algorithm performs well on synthetic data.

References

[1] A. Zeira and B. Friedlander, "Direction finding with time-varying arrays," *IEEE Transactions on Signal Processing*, vol. 43, pp. 927–937, April 1995.

[2] B. Friedlander and A. Zeira, "Eigenstructure-based algorithms for direction finding with time-varying arrays," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, pp. 689–701, April 1996.

[3] J. Sheinvald, M. Wax, and A. J. Weiss, "Localization of multiple sources with moving arrays," in *Proc. ICASSP '97*, pp. 3521–3524, 1997.

[4] F. C. Robey, "A covariance modeling approach to adaptive beamforming and detection," Tech. Rep. 918, MIT Lincoln Laboratory, July 1991.

[5] A. D. Lanterman, "Statistical imaging in radio astronomy via an expectation-maximization algorithm for structured covariance estimation," in *Statistical Methods in Imaging: In Medicine, Optics, and Communication* (J. A. O'Sullivan, ed.), Springer-Verlag, 2000. in press.

[6] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, pp. 963–974, Sept. 1982.

[7] T. A. Barton, *Covariance Estimation for Multidimensional Data.* PhD thesis, Washington University in St. Louis, 1993.

[8] D. L. Snyder, J. A. O'Sullivan, and M. I. Miller, "The use of maximum likelihood estimation for forming images of diffuse radar targets from delay-doppler data," *IEEE trans. on Information Theory*, vol. 35, pp. 536–548, May 1989.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society*, vol. 39, pp. 1–37, 1977.

LIST OF AUTHORS

Abed-Meraim, K. 193
Abou-Faycal, I. 337
Ahmed, M. 755
Akansu, A.N. 963
Al-Dhahir, N. 680
Aleshnikov, I. 1031
Alesiani, F. 297
Al-Naffouri, T.Y. 1175
Alouini, M.-S. 42, 347, 367
Altman, E. 537
Anantharam, V. 307
Ananthaswamy, G. 377
Anigstein, P. 307
Badowski, G. 528
Balakrishnan, J. 1214
Bandyopadhyay, S. 1305
Barg, A. 81
Barton, R.J. 1
Bayoumi, M.M. 1155, 1165
Berger, T. 700
Bertsekas, D.P. 872
Bi, C. 225
Biglieri, E. 297
Black, C.K. 896
Blake, I.F. 238
Blight, D. 1266
Boariu, A. 997
Borkar, V.S. 849
Bourdon, S.E. 1145
Brehler, M. 1067
Burkykh, S. 193
Burrell, A. 577
Buzzi, S. 725, 816
Cabral, H.A. 69
Caire, G. 715
Capone, J.M. 133
Chan, A.M. 668
Charpin, P. 1029
Cheng, J. 700
Cheng, P. 926
Chindapol, A. 975
Choi, H. 1314
Choi, H.-A. 1314
Chong, E.K.P. 113, 287
Chow, P.L. 412
Chowdhury, S. 1224
Chugg, K.M. 689, 959
Corcoran, J.N. 1116
Coskun, O. 689
Costello, Jr., D.J. 69, 468
Dahleh, M.A. 916
Damen, M.O. 193
Das, D. 1194
de Farias, D.P. 855
del Angel, G. 327
Dube, P. 537
Dumer, I. 71
Dutta, R. 1284
East, J. 785
Ephremides, A. 103, 836
Erez, E. 498
Fan, J.L. 955
Feder, M. 498
Fekri, F. 957
Feng, G.-L. 593
Feron, E. 449, 916
Fine, T.L. 327
Fitz, M. 985
Fitz, M.P. 965, 995, 1057
Flanagan, J.L. 1244
Francis, B.A. 143
Frazzoli, E. 916
Frey, B.J. 62
Fuhrmann, D.R. 648
Gajic, Z. 393
Ganesan, A. 191
Garg, A. 1256
Gelfand, S.B. 357
Gentry, S. 449
Georghiades, C.N. 636, 886, 967
Giannakis, G.B. 52, 1076
Goeckel, D.L. 377
Goldsmith, A. 745
Goldstein, J.S. 1224
Goodwine, B. 906
Gosavi, A. 865
Grossglauser, M. 735
Guess, T. 795, 806
Gulliver, T.A. 258
Gupta, R. 785
Guruswami, V. 603
Gutierrez, F.J. 1294
Hagenauer, J. 1234
Hassibi, B. 1047
Hastad, J. 603
Heath, Jr., R.W. 212, 1086
Henderson, S.G. 1125
Hero, A. 785
Hochwald, B.M. 1047
Høst-Madsen, A. 32
Hottinen, A. 997
Hou, J. 266
Howitt, I. 847
Hui, J.Y. 222, 225
Huisinga, W. 1106
Ishii, H. 143
Ji, T. 961
Johansson, M. 508
Johnson, Jr., C.R. 1214
Jost, R. 133
Kalyanasundaram, S. 113
Khammash, M.H. 163
Khasminskii, R. 412
Kieffer, J. 480
Kim, H.S. 557
Knap, S. 1155, 1165

Koetter, R. 625
Koksal, A.E. 872
Kontoyiannis, I. 461
Kumar, P.V. 845, 1031
Kurzhanski, A.B. 403
Kwong, R.H. 1135
Lafortune, S. 785
Lam, W.H. 387
Laneman, J.N. 775
Lanterman, A.D. 648
Lanzetta, G. 69
Lau, W.C. 347
LaValle, S.M. 926
Lawford, M. 1145
Le Martret, C. 52
Lee, E.B. 439
Li, H. 1000
Li, J. 1000
Li, Y. 836
Liang, P. 785
Liberty, S.R. 422
Lim, J.-H. 357
Liu, Y. 995
Liu, Z. 1076
Lops, M. 816
Lott, C.G. 123
Lu, H.F.F. 845
Lynch, K.M. 896
Madhow, U. 337
Mandjes, M. 567
Mannor, S. 470
McCloud, M.L. 1067
McLaughlin, S.W. 957
Médard, M. 337, 583
Mersereau, R.M. 957
Milstein, L.B. 266
Mina, M. 1274
Mitra, U. 22
Mitter, S. 83, 93, 182, 451
Morishima, S. 1247
Mou, L. 422
Müller, R.R. 1096
Murphy, R.J. 646
Naguib, A.F. 680
Nassar, C.R. 203
Neely, M.J. 547
Nguyen, G.D. 103
Ni, J. 1
O'Sullivan, J.A. 646
Oklander, B. 488
Olshevsky, V. 613
Onggosanusi, E.N. 228
Paganini, F. 153
Papadopoulos, H. 1184
Papantoni-Kazakos, P. 577
Pauciullo, A. 816
Paulraj, A. 212, 1086
Pavlović, V. 1256
Pham, K.D. 422
Pi, Z. 22
Pless, V. 1010
Poor, H.V. 725
Pottie, G. 755
Qi, X. 163
Qian, L. 393
Raghothaman, B. 997
Rankin, D. 258
Rehg, J.M. 1256
Ricci, G. 816
Richa, A.W. 1305
Richie, J. 847
Rieken, D.W. 648
Ritcey, J.A. 975
Rohrs, C.E. 547
Rosenthal, J. 248
Rouskas, G.N. 1284
Saengudomlert, P. 583
Sahai, A. 83
Sain, M.K. 422
Salapaka, M.V. 163
Sandhu, S. 1086
Saniee, I. 567
Sasaki, G.H. 1266
Sason, I. 276
Sayed, A. 1175
Sayeed, A.M. 191, 228, 705
Schafer, R.W. 957
Schniter, P. 1204
Schütte, C. 1106
Seidl, N. 847
Sen, A. 133, 1305
Shabunov, K. 71
Shamai, S. 715
Shamir, G.I. 468
Shattil, S. 203
Shen, B.H. 1305
Shen, Z. 926
Shepard, T. 1155, 1165
Shi, Y.-Q. 945
Shokrollahi, A. 613
Shroff, N.B. 113, 557
Shum, K. 1031
Siegel, P.H. 266
Simon, M.K. 42, 347
Siwamogsatham, S. 1057
Snyder, D.L. 646
Somani, A.K. 1274
Spasojević, P. 636, 886
Stamoulis, A. 1076
Stark, W. 785
Stark, W.E. 961
Stichtenoth, H. 1031
Stolyar, A.L. 567
Su, C.-F. 1266
Subramaniam, S. 1314
Sudan, M. 603
Sun, H. 225
Sun, J. 965
Sun, Y. 826
Surijan, E. 488

Sutskover, I. 488
Takeshita, O.Y. 965, 995
Taricco, G. 297
Tatikonda, S. 93, 182, 451
Telatar, E. 276
Teneketzis, D. 123, 785
Thiennviboon, P. 959
Tilbury, D.M. 172
Tirkkonen, O. 997
Tong, L. 765
Toumpis, S. 745
Tse, D. 735
Tu, Y.-S. 755
Tüchler, M. 1234
Tulino, A.M. 11
Tweedie, R.L. 1116, 1125
Ulukus, S. 317
Urbanke, R. 276
Uysal, M. 967
Van Roy, B. 855
Van Veen, B.D. 228
Varaiya, P. 403
Varanasi, M.K. 795, 1067, 1194
Vardy, A. 625
Varvarigos, E.A. 1294
Vassiliadis, S. 1294
Venkatesh, S. 449
Verdú, S. 11
Viswanathan, H. 658
Viterbo, E. 297
Vontobel, P.O. 248
Walker, J.L. 1019
Wang, H. 785
Wang, X. 32
Wasserman, K.M. 518
Wei, X. 963
Wei, Y. 906
Weissman, T. 470
Wieselthier, J.E. 103
Wonham, W.M. 1135, 1145
Wood, J.A. 1038
Wornell, G.W. 668, 775
Worthen, A. 785
Xiao, L. 508
Yang, E.-H. 480
Yang, H. 367
Yates, R.D. 317
Yaz, E.E. 431
Yaz, Y.I. 431
Yin, G. 528
Yook, J.K. 172
You, K.H. 439
Zad, S.H. 1135
Žefran, M. 936
Zekavat, S.A. 203
Zémor, G. 81
Zhang, D. 518
Zhang, J. 287
Zhang, Q. 528
Zhang, X. 945, 985
Zhao, Q. 765
Zheng, Z.H. 387
Zhou, S. 52
Ziv, J. 488
Zoltowski, M.D. 1224
Zuckerman, D. 603

3 0112 050707642